AD_____

Award Number: DAMD17-98-1-8323

TITLE: Deriving Structures for Lead Drug Discovery from Cell-Line Screens

PRINCIPAL INVESTIGATOR: Robert L. Jernigan, Ph.D.

David G. Corell, Ph.D.

CONTRACTING ORGANIZATION: Department of Health and Human Services
National Cancer Institute
Bethesda, Maryland 20892

REPORT DATE: September 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1999	Annual (1 Sep 98 - 31 Aug 99)

4. TITLE AND SUBTITLE
Deriving Structures for Lead Drug Discovery from Cell-Line Screens

5. FUNDING NUMBERS
DAMD17-98-1-8323

6. AUTHOR(S)
Robert L. Jernigan, Ph.D.
David G. Corell, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Health and Human Services
National Cancer Institute
Bethesda, Maryland 20892
E-MAIL:
jernigan@structure.nci.nih.gov

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

New calculational methods have been developed and applied towards the discovery of novel drugs for the treatment of breast cancer. The analysis is based on publically available tumor screening data generated in the National Cancer Institute's anticancer drug screen. The growth inhibitory potency of 122 anticancer agents tested in the NCI screen have been analyzed using methods of singular value decomposition (SVD). The analysis yielded clusters of compounds with similar activity patterns and compared these results with their putative mechanism of action (MOA). Clustering according to each compound's vector of screening activity segregated compounds into two groups, clearly discernible on the basis of pattern similarities in their potency. The first group includes compounds that act as DNA-damaging agents while the second group includes compounds that act as inhibitors of biosynthetic enzymes or mitosis. Additional analysis of an expanded set of tested compounds finds that a significant statistical correlation can be found between compounds that have similar functions and compounds with structural similarities. These results represent the first evidence for a strong correlation between cancer screening data and structure.These results provide a basis for further explorations of relationships between structural modalities of potential anticancer agents and measurements of cellular toxicity.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast Cancer	61
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

D.C. Where copyrighted material is quoted, permission has been obtained to use such material.

D.C Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

D.C Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

David G. Covell 9-27-99

PI - Signature Date

TABLE OF CONTENTS

III. Introduction

The overall goal of the proposed research is to apply new calculational methods towards the discovery and development of novel drugs for the treatment of breast cancer. The research is focused on analyses of publically available databases of compounds that have been screened against selected panels of tumor cell lines. These databases tabulate the ability of candidate compounds to limit the growth of tumor cells. The project will conduct a detailed analyses of these data to determine which classes of compounds are most effective against selected tumor cell panels. Calculational methods initially focus on cluster analysis, directed at the identification of compounds that elicit the most similar responses across tumor cell types. This analysis is conducted in parallel with a similar analysis based on common structural features between tested compounds. The results from each of these analyses provide a means to relate functional properties (i.e. inhibition of cell growth) to structural features of tested compounds, thereby creating a foundation for comparisons with similar databases of information. The project will initially focus on the extraction of compounds that represent completely novel agents, agents with structurally and functionally unique descriptors, from compounds similar to previously tested agents. The intention will be to contrast results between various tumor panels, and from this differential activity identify compounds most active against tumor cells within the breast panel. A secondary component of this analysis will be the characterization of activities within and between different tumor cell panels. These differences can then be analyzed in the context of genomic expression profiles, as they become available.

IV. Body

We are currently on schedule with respect to our proposed statement of work (SOW). Summary statements for each of these Tasks are provided below. In general, this research project is proceeding according to the proposed work scope. The research listed as Task 1 has resulted in a recently submitted manuscript for a peer reviewed journal (see Appendix). A second manuscript, documenting our results for Task 2, is currently nearing completion. Task 3 extends into the next year, however, it is included in the BODY description because we are beginning this phase of the project and have additional information to report.

A. Task 1: Months 1-6: Tabulation of Cell-Line Screening Data. Develop computer programs to analyze these data

The NCI has currently tested nearly 30,000 synthetic products against a panel of 60 selected tumor cell lines. Endpoints of the NCI's 60 tumor cell line screen include the growth inhibitory activity ($\log(GI_{50})$) of each compound, expressed as the drug concentration required to inhibit cell growth by fifty percent compared to an untreated control. $\text{Log}(GI_{50})$ values for a single compound across all 60 tumor cell lines provide an activity pattern which can be compared to patterns for other tested compounds. A systematic analysis of this data began with a variety of statistical tests to 1) evaluate the quality of this data, 2) identify and resolve data entry errors versus real errors, and 3) reduce this dataset to data of the highest quality for further evaluation. Our initial research effort focused on cell screening data for a set of standard anticancer agents currently used in the clinic. This initial set comprised 122 compounds and the screening data against these compounds should provide a strong reference point for comparisons to compounds in larger test sets. The manuscript included in the Appendix provides a detailed description of our efforts to organize the cell-screening data and develop computer programs to analyze the data from the 122 standard agents. Without major redundancy, we summaize these findings as reported in the Abstract.

Computational programs were successfully completed for statistical analysis of the cell-screening data. Data analysis with these computational tools

found that the 122 standard agents could be divided into 25 statistically distinct groups. Within these groups, 8 groups include structurally diverse compounds with reactive functionalities that act as DNA-damaging agents while 17 groups include compounds that inhibit nucleic acid biosynthesis and mitosis. These data provided a reference response for agents against each of these 60 tumor cell lines. A companion analysis directed at clustering each of the cell types by their response to the 122 anticancer agents divided the 60 tumor cell types into 21 groups. The strongest within-panel groupings were found for the renal, leukemia and ovarian cell panels. A coherent set of cells within the breast cancer cell lines was found, however, a stronger coherence was found between breast, prostate and colon cell lines.

B. Task 2: Months 6-10: Group active compounds and find substructure similarities identified in Task 1.

The functional clusters obtained within Task 1 provide a means to compare activity patterns with substructure features. A diverse set of structural features are observed for compounds within these functional groups, with frequent occurrences of strong within-group structural similarities (see Figure 1 of the attached manuscript). This analysis provides a baseline for further comparisons within the larger set of tested compounds. Using standard 1D substructure similarity searches, based largely on SMILES-based descriptors, a list of 272 compounds are found within the NCI's database of tested compounds with strong structural similarity to the set of 122 standard anticancer agents. The tumor cell-line screening data for these compounds has been examined using the tools developed in Task 1. It is noteworthy that the majority of compounds in this expanded dataset share structural similarities with the antimitotic agents in the 122 standard agents thought to be active against breast cancer. Thus this set, totaling 394 compounds, provides a rich starting point for closer examination of relationships between potency and cell specificity for these agents.

Our computational tools have been successfully used to identify the greatest overlaps between statistical clusters based on functional data (i.e. patterns in the cell-line screen) and clusters based on structural features. The goal here is to reorganize sets of functional and structural clusters so as to maximize compound membership within each cluster. Analysis of the

$\log(GI_{50})$ data finds that the 394 compounds represent 60 functional clusters. Using the methods developed in Task 1, the distances between each set of NSC compounds, as determined by their patterns in the 60 tumor cell screen, are calculated and displayed spectrally in Figure A. Dark blue indicates compounds with the closest response patterns, green, intermediate and red the most distant response patterns. In this figure the compounds within the same cluster appear together and the clusters have been ordered such that the most similar patterns appear adjacent to each other. In this figure, alkylating agents appear in the top left-most portion of the figure, agents that act as antimitotics are in the middle and agents that act as inhibitors of nucleotide biosynthesis appear at the lower right portion of the figure. Although each of the 60 clusters will not be described in detail here, individual clusters appear in this figure as blocks of dark blue (i.e. close response neighbors) along the diagonal.

Our analysis of the structural similarities of these compounds finds that there are 150 structural classes within this set of 394 compounds. The 60 functional clusters and the 150 structural clusters are further examined for clusters that share the greatest number of compounds. The intention here is to identify those compounds that share both structural and functional similarity. These compounds provide a basis for assignment of a structural pharmacophore. This pharmacophore can be used to identify additional compounds that might exist within larger databases. More importantly, however, this pharmacophore can be contrasted with compounds that share this structural feature, but are associated with a completely different functional response. Such differences can be used to suggest distinctive substructural features that may be related to each different functional response. Figure B displays those compounds that appear jointly in a structural and functional cluster. A one-to-one correspondence between functional and structural clusters would appear as a diagonal line in this plot. Since there is not a one-to-one correspondence (i.e. there are an unequal number of functional and structural clusters), a near-diagonal location indicates such a correspondence. The compounds in Figures A and B are ordered identically, so that a visual comparison can be made between functional classes (Figure A) and structural classes (Figure B). Examination of the compounds at various diagonal positions of this plot reveals a strong correspondence between structural features and their putative mechanism of action. Thus alkylat-

ing agents, which appear as the first 100 compounds in Figures A and B, share a variety of functional features, most notably chloro-ethylating groups. Compounds appearing along the diagonal at the middle and lower right portions are less structurally diverse, and by examination represent structurally coherent groups of compounds.

In summary, this effort has produced a tool for careful examination of cluster overlap for any chosen clustering scheme. As a result, a consensus cluster order can be achieved, based on multiple clustering schemes and diverse data (structural versus functional). These consensus clusters are currently being examined in an effort to generate the most dominant structural features shared by compounds within a given cluster. As an example, the subset of compounds derived from the natural product family of trichothecenes (otherwise referred to as VERRUCARIN), has been found to produce a distinctive response within the cell screen. This family of compounds represents a case with unique structural and functional features. A detailed examination of this unique response in currently underway. The above research results are currently being developed in a manuscript, estimated to be completed during the last quarter of this calendar year.

C. Task 3: Months 11-18: Develop initial consensus models of drugs active against breast cancer cell-line.

As indicated in the description of Task 2, the 394 compounds derived from the standard 122 agents and analyzed with our new computational tools includes a large group of compounds thought to act as antimitotics. Our results from Tasks 1 and 2 have led to the development of tools for examining correlations between structural and functional clusters. This information, coupled with additionally available tools for structural analysis, is being applied to this dataset. The immediate goal is to clearly identify a structural pharmacophore that spans one or more of the consensus clusters for the antimitotic agents. This pharmacophore (or pharmacophores) will then be used to scan the entice set of compounds listed in the NCI's database. Currently this database includes nearly 400,000 structures, with cell screening data available for only 10 percent of these compounds. The intention here is to identify a population of untested compounds with strong structural similarities within

the set of 394 compounds. From this set of compounds, additional screening tests will be requested. The screening effort will take place simultaneously with the modeling studies of ligand docking to selected available crystallographic structures. Although the currently available crystal structures of tubulin are the initial choice for docking studies, we have additional evidence that selected antimitotic compounds may also preferentially alkylate DNA.

Figure A: Functional Clusters, 394 Compounds

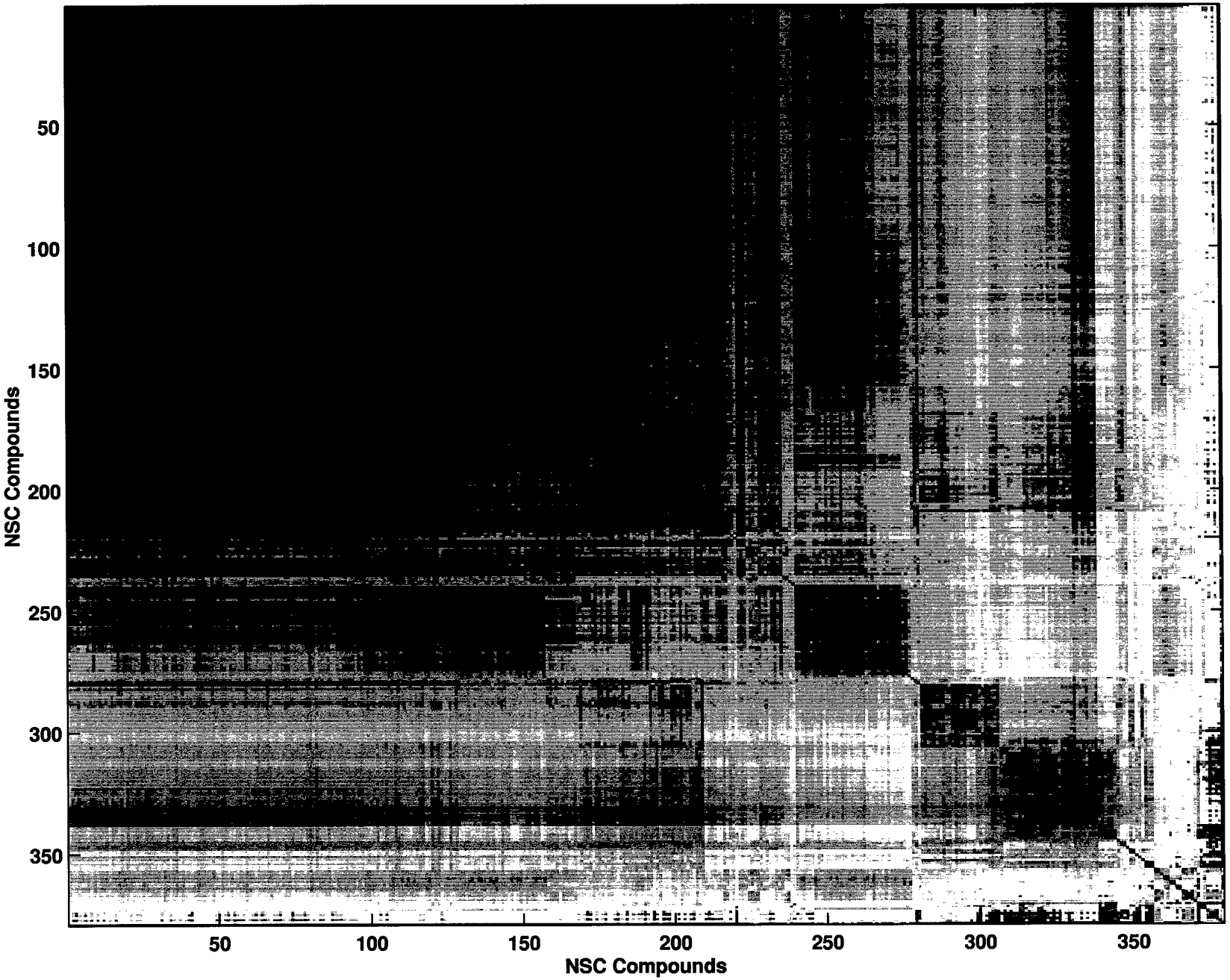
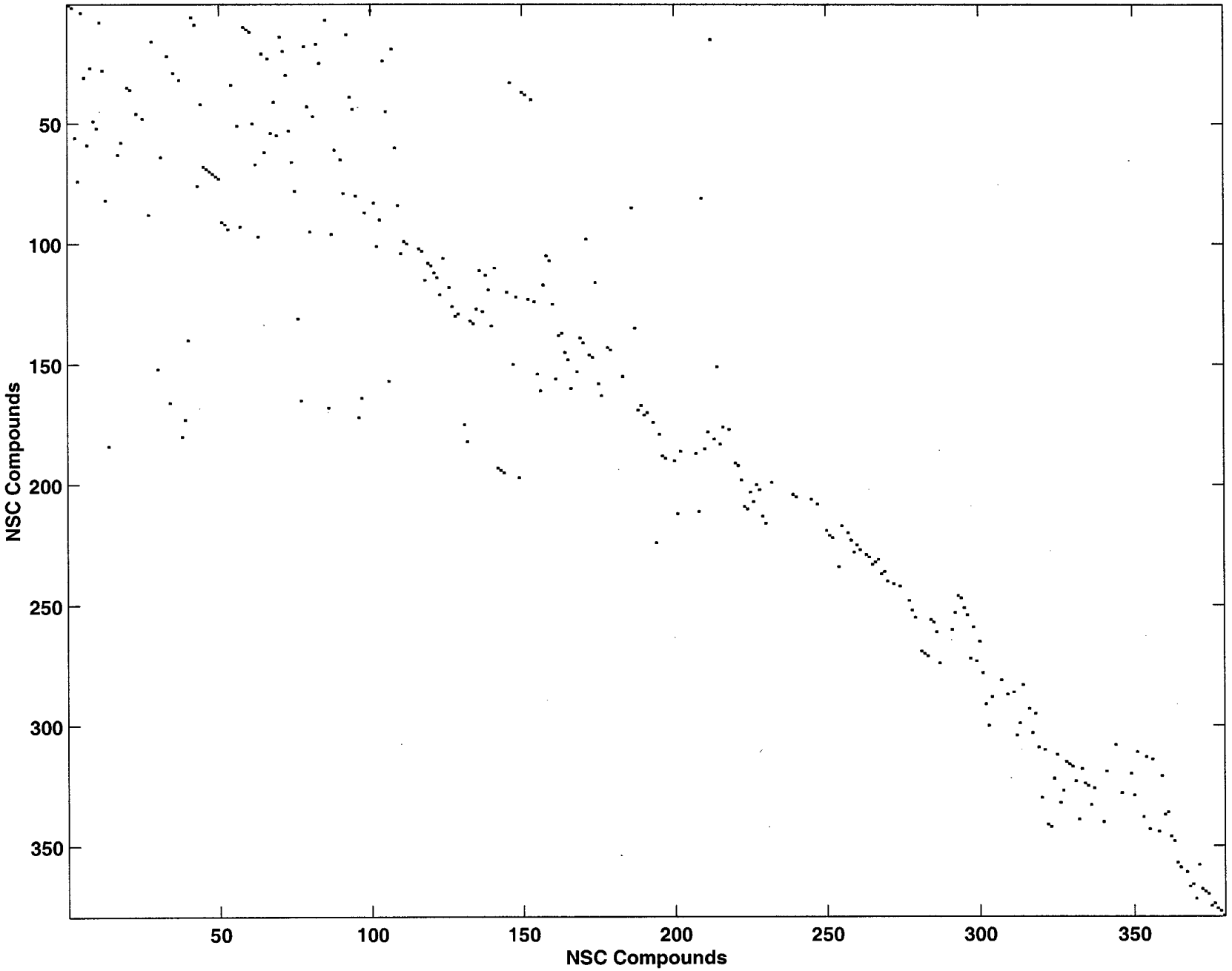


Figure B: Structural Clusters, 394 Compounds



VII. Key Research Accomplishments

- High quality refinement of NCI tumor cell screening data
- Development and application of statistical clustering software for analysis of tumor cell screening data
- Identification of compound clusters based on similarities in activity patterns in the cell screen
- Integration of clustering results based on screening dataset and clustering results based on substructure identification
- Implementation of tools for efficient scanning of entire NCI database of small synthetic molecules

VIII. Reportable Outcomes

The results obtained in Task 1 were presented at two national meetings for the Biophysical Society and the Protein Society. Both of these meetings led to scientific communications with others in the field of breast cancer research. The manuscript reporting our results from Task 1 is sttached in the Appendix. This paper has been submitted the the journal, Cancer Research.

IX. Conclusions

The current research results provide a systematic analysis of the currently available data from the NCI's cancer screening project. Our efforts, to date, have established a link between the activity patterns from the tumor cell screen and the common structural features of the tested compounds. Our analysis has identified a subset of antimitotic agents for which cell screening data does not yet exist. Efforts to examine these compounds for common substructures are currently underway.

X. References

Most of the references appropriate for this report are found in the manuscript provided in the Appendix

XI. Appendices

Characterization of Anticancer Agents by Their Growth-Inhibitory Activity and Relationships to Mechanism of Action and Structure

Ozlem Keskin[1], Ivet Bahar[1], Robert L. Jernigan[2], Timothy G. Myers[3], John A. Beutler[4], Robert H. Shoemaker[3], Edward A. Sausville[3] and David G. Covell[2]

[1]*Chemical Engineering Department and Polymer Research Center, Bogazici University, TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey*

[2]*Molecular Structure Section, Laboratory of Experimental and Computational Biology, NCI, NIH, SAIC, Frederick MD 21702, and Bethesda, MD 20892 USA*

[3]*Developmental Therapeutics Program, Divison of Cancer Treatment and Diagnosis NCI, NIH, Frederick MD 21702, and Bethesda, MD 20892 USA*

[4]*Laboratory of Drug Discovery Research and Development, DTP, DCTDC, NCI, SAIC, Frederick, MD 21702*

Keywords: tumor cell-line screen, SVD, clustering behavior

Running Title: SVD Analysis of Cell Screening Data

Abstract

An analysis of the growth inhibitory potency of 122 anticancer agents available from the National Cancer Institute (NCI) anticancer drug screen is presented. Methods of singular value decomposition (SVD) were applied to determine the matrix of distances between all compounds. These SVD-derived dissimilarity distances were used to cluster compounds that exhibit similar tumor growth inhibitory activity patterns against 60 human cancer cell lines. Cluster analysis divides the 122 standard agents into 25 statistically distinct groups. The first 8 groups include structurally diverse compounds with reactive functionalities that act as DNA-damaging agents while the remaining 17 groups include compounds that inhibit nucleic acid biosynthesis and mitosis. Examination of the average activity patterns across the 60 tumor cell lines reveals unique 'fingerprints' associated with each group. A diverse set of structural features are observed for compounds within these groups, with frequent occurrences of strong within-group structural similarities. Clustering of cell types by their response to the 122 anticancer agents divides the 60 cell types into 21 groups. The strongest within-panel groupings were found for the renal, leukemia and ovarian cell panels. These results contribute to the basis for comparisons between log(*GI(50)*) screening patterns of the 122 anticancer agents and additional tested compounds. (IRSP, SAIC Frederick, NCI-FCRDC; Funded in part by NO1-56000, DAMD17-98-1-8323 and the ARMY Breast Cancer Research Project)

Introduction

Development of high-throughput screening technologies in drug discovery has led to dramatic increases in the diversity of compounds that can be tested [1-3] and in the types of targets available for testing [4-11]. Accompanying these advances has been the development of a diverse collection of general approaches for mining the large quantity of data generated by these systems [12-20]. Database-related, information-intensive drug discovery efforts [21] are showing promise in revealing relationships between drug screening profiles and potential therapeutic targets. Extending these efforts by further exploration of relationships between screening profiles and chemical structures may enhance the discovery of novel chemotherapeutic agents.

In this paper we reexamine the publicly available data from the cancer drug discovery program at the National Cancer Institute (NCI). Our goal is to systematically analyze the relationship between 1) the growth inhibitory activities for a set of anticancer agents from the panel of 60 tumor cell lines, 2) the structural features of the tested agents and 3) their apparent mechanism of growth inhibitory action (MOA). Based on the hypothesis that selective *in vitro* activity of a compound against cancer cell lines might be predictive of its activity against the corresponding specific type of human tumor, the NCI has developed and made available, results of primary drug screens against 60 different human cancer cell lines (http://dtp.nci.nih.gov). Among other endpoints available in the NCI's database, the growth inhibitory activity of each compound, expressed as the drug concentration (GI_{50}) required to inhibit tumor cell growth by fifty percent compared to an untreated cell was selected for analysis. Log(GI_{50}) values for a given compound across all tumor cell lines provide its activity pattern for comparison to patterns from other tested compounds. Similarities in patterns of *in vitro* inhibitory activity have been shown to be related to MOAs, modes of resistance and molecular structure [4, 16, 19-24]. To date, the NCI has screened over 70,000 chemical compounds and a similar number of natural product extracts against a panel of 60 different tumor cell lines.

Several algorithms have previously been applied to analyze activity patterns. These algorithms utilize, in various ways, the tools of multivariate statistical

clustering [25]. As an example, the internet accessible program COMPARE [22, 23] uses Pearson correlation coefficients (PCCs) to extract compounds with screening patterns similar to a 'seed' compound. Applications of back-propagation neural networks [26] and Kohonen self-organizing maps [27] have demonstrated varying success when predicting MOA, and grouping compounds based on similar activity patterns. These methods also complement the COMPARE program by identifying clusters of 'seed' compounds, thus addressing the important question of whether a 'seed' compound appears on the lists of highly correlated activity patterns for all other 'seeds' in the dataset. Statistical and artificial intelligence techniques, including principal component analysis, hierarchical cluster analysis, stepwise linear regression and multidimensional scaling, have begun to be applied to the NCI's screening data [28, 29].

Structurally similar compounds can have similar physicochemical properties and thus are thought to have similar biological activities, consistent with the similarity property principle [30]. For example, a dramatic coherence between molecular structures and activity patterns was observed for 112 ellipticine analogs [19]. Detailed crystallographic and NMR studies further support the similarity property concept by demonstrating that ligand-receptor interactions are characterized by complementary shapes and chemical characteristics [31-34]. Cell-based screening assays represent a complex array of interactions that is monitored as cell growth or killing (e.g. log (GI_{50})). Differential activity patterns in these measurements can result from the activity of compounds that interact well, poorly, or not at all, with one or many targets within the panel of cell types. Earlier attempts to establish correspondences between activity patterns, MOAs and chemical structure found general clustering a) for compounds of similar chemical structure, and b) for compounds classified as having a similar mechanism of action (MOA), yet having diverse chemical structures [28]. Distant clustering was also found for compounds similar in chemical structure but having different MOAs [28]. Earlier studies by Paull et al. [23, 20] demonstrated that anticancer agents having similar functional groups (e.g. chloroethylating agents, platinum analogs and nitrosoureas) produce similar activity patterns in cell-based screens. However, there are some compounds that display a relatively strong structural similarity, and yet exhibit drastically different activity patterns. Alternatively, compounds with

similar activity patterns can have little structural correspondence to one another.

The present analysis identifies clusters of anticancer compounds based on their log(GI_{50}) activity patterns in NCI's data for 60 tumor cell lines. The analysis is performed on the set of 122 standard anticancer agents available in the NCI's Developmental Therapeutic Program's database. Here we adopt singular value decomposition (SVD) [35-39] and hierarchical clustering methods [40] to cluster the chemotherapeutic agents. Compounds clustered with these methods are to be compared by their assigned MOAs and their structural similarities.

Methods

Variance-based measures of similarity rely on the spread in a dataset to determine membership within a cluster. Principal component analysis (PCA), SVD, D-optimal design and k-nearest neighbor clustering are commonly used as variance-based methods. These have as their overall goal, the minimization of noise-to-signal ratio [41]. The SVD approach has been shown to be a powerful method to filter noise and enhance the information content of the original data [35-38]. Similarly to PCA, SVD defines rotation of axes (principal components) so that columns in the data matrix maximize their standard deviation with respect to other columns in the dataset. This transformation yields a new space where the columns of data exhibit maximum variance (i.e. minimum correlation) with respect to one other. The original data can be re-expressed approximately as a linear combination of a few dominant principal components. This new space, referred to as the SVD space, has previously been effectively used, for example, to classify words within texts [36], and protein structures with respect to their amino acid composition [42].

SVD analysis is used here to classify anticancer agents by examining their log(GI_{50}) values in the 60-dimensional space of the cancer cell lines. This space is transformed into an SVD space, where the anticancer agents are represented by activity arrays emphasizing their differences. The compounds are clustered on the basis of their pairwise distances in the SVD space, by using hierarchical clustering algorithms [40]. The calculations discussed below have

been coded into a Fortran program, which is available upon request. Many of these calculations can also be completed using the SAS library of utilities.

In general, the SVD of a given matrix **A** yields three matrices **U**, Λ, and **V** which comprise (i) the singular values λ_i of **A**, organized in ascending order in the diagonal matrix Λ, (ii) the orthonormal transformation matrix **U** ensuring the passage between the original coordinate frame and the SVD frame, and (iii) the normalized representation, $\mathbf{V}^T$, of the original matrix in the SVD space. **A** is expressed as the product as these matrices

$$\mathbf{A}_{mxn} = \mathbf{U}_{mxm}\ \Lambda_{mxm}\ \mathbf{V}^T_{mxn} \qquad (1)$$

where the subscripts denote the dimensions of the two dimensional matrices, and the superscript T indicates the transpose. In general, the columns of **A** each represent a given quantity (here anticancer agents) characterized by m properties (activity patterns for 60 cell lines); whereas those of the product $\Lambda\mathbf{V}^T$ are the same quantities expressed in the SVD frame which best describes the similarities/ differences between these quantities on the basis of their n properties. In the present application of the SVD method to anticancer compound screening data, each column of **A**, conveniently denoted as $\mathbf{a}_i$, is a 60-dimensional vector describing the activity pattern of a given drug i ($1 \leq i \leq 122$), expressed in terms of the logGI_{50} values observed against the 60 tumor cell lines. Therefore the SVD of a 60 x 122 matrix is performed, using the dataset of n = 122 anticancer agents screened against m = 60 cell lines. The ijth element of this matrix (or the jth element $[\mathbf{a}_i]_j$ of the ith column $\mathbf{a}_i$) is defined as

$$A_{ij} = [\mathbf{a}_i]_j = \Delta x_{ij} - \langle \Delta x \rangle_j \qquad (2)$$

where x_{ij} is the logarithmic concentration of the anticancer agent i for the inhibition of the growth of the jth cell line by fifty percent, designated as $[\log GI_{50}]_{ij}$, and Δx_{ij} is the differential change in this value relative to the average cytotoxic potency, $\langle \log GI_{50} \rangle_i$, of the particular agent over the entire panel of cells, i.e.

$$\Delta x_{ij} = [\log GI_{50}]_{ij} - \langle \log GI_{50} \rangle_i \qquad (3)$$

and finally $\langle \Delta x \rangle_j$ is the average of Δx_{ij} over all agents for the particular cell line. Subtraction of $\langle \Delta x \rangle_j$ in eq 2 eliminates the differences arising from the generic characteristics of the particular cell lines, and permits us to emphasize more clearly the differences among activity patterns of the anticancer agents. The activity fluctuation pattern of the ith agent in the SVD space is represented by the ith column $\mathbf{v}^T_i$ of $\mathbf{V}^T$ pre-multiplied by Λ, and designated as $\mathbf{a}_i^* = \Lambda \mathbf{v}^T_i$ such that the SVD distance between agents i and j is

$$d_{ij} = [(\mathbf{a}_i^* - \mathbf{a}_j^*) \bullet (\mathbf{a}_i^* - \mathbf{a}_j^*)]^{1/2} = [(\Lambda \mathbf{v}^T_i - \Lambda \mathbf{v}^T_j) \bullet (\Lambda \mathbf{v}^T_i - \Lambda \mathbf{v}^T_j)]^{1/2} \qquad (4)$$

The above distances constitute the basic measure for clustering the anticancer agents into groups in the present SVD analysis. The analyzed set includes 122 compounds with six putative MOAs: 35 alkylating agents, 24 antimitotic agents, 16 topoisomerase I inhibitors, 19 topoisomerase II inhibitors, 16 RNA/DNA antimetabolites, and 13 DNA antimetabolites.

Results

The results of clustering compounds according to their pairwise SVD distances are listed in Table 1. Clusters obtained from pairwise distances place compounds with the most similar activity patterns adjacent to one another. Using this approach, clusters are ordered such that compounds with the greatest and least similarities in their SVD distances are presented first and last, respectively, in Table 1. Figure 1 displays the 2D structures of the compounds within each cluster.

Statistical clustering of these patterns was obtained using the SAS/STAT clustering algorithms. The cubic clustering criterion (CCC) was selected to determine cluster membership. This criterion estimates the number of clusters based on minimizing the within cluster sum of squares. The CCC calculation generates a rough approximation to a 'goodness of fit' measure under the null hypothesis that the data are sampled from a uniform distribution on a hyperbox (p-dimensional right parallelpiped). A t-test statistic with one degree of freedom ($t=3.078$, $p<0.05$, $n=1$) is generated for testing the null hypothesis that a compound's SVD distance pattern is not different from a given cluster (i.e. cannot be excluded from the cluster). This method has been shown to help

determine cluster number for both univariate and multivariate data with small sample sizes (n~20). See SAS Technical Report A-108 for additional details.

The results of this analysis find that the 122 standard agents can be clustered into 25 groups, labeled GROUPS 1-25, and listed in Table 1. Fifteen of these groups have at least 2 members, while the final 10 groups consist of a single agent. Figure 1 displays the molecular structures of these compounds, ordered according to the GROUPS 1-25 in Table 1. The list of compounds in each group in Table 1 includes their putative MOAs and characteristic structural/functional groups. Multiple compounds within each group cannot be further subdivided on the basis of their log(GI_{50}) patterns. However, structural similarities within clusters can be easily found by inspection of Figure 1.

Group 1 is composed of 38 compounds consisting predominantly of alkylating agents (23 compounds), topoisomerase II inhibitors (9 compounds), DNA antimetabolites (5 compounds) and a single RNA/DNA antimetabolite. Alkylating agents are antitumor drugs that act through covalent binding of their alkyl groups to cellular molecules [43, 44]. Many of these are proposed to attack the N-7 or O-6 atoms on guanine in the DNA major groove, and to cross-link DNA strands [43, 44]. Cross-linked products are removed by an alkyltransferase DNA repair enzyme, via a repair mechanism known to be deficient in certain tumors. The first two members of this group are compounds bearing two or more aziridine or oxirane groups (296934 and 182986). These are analogs of the putative closed-ring intermediates of the nitrogen mustards, but are believed to be less reactive [43]. Three of the five platinum containing compounds are found next within this group (119875, 256927 and 241240). The next set of compounds in this group is composed of alkyl alkane sulfonates (329680, 102627, 750, 348948 and 338947). Busulfan (750) has been shown to attack the N-7 atom of guanine, but its ability to cross-link DNA is not certain. Pyrazoloimidazole (51143) and guanazole (1895) appear next, and are highly reactive DNA antimetabolites with nitrogen containing ring structures. The prodrug ftorafur (148958) appears next. The remaining members of Group 1 fall into two structural classes: the first composed of nitrosoureas, either alone, or in combination with nitrogen mustards or guanidine groups (32065, 8806, 3088, 25154, 73754, 353451, 409962, 171112, 95441, 178248, 95466, 79037, 95678, 107392, and 167780), and the second composed of anthracyclines, anthracenediones and podophyllotoxins (308847, 142892,

366140, 349174, 355644, 164011. 123127, 82151, 267469 and 141540). The nitrosourea compounds bearing both chloroalkylating and carbamoylating (carbamoyl: -R-N-C=O) groups can produce interstrand cross-links in DNA by preferentially attacking the O-6 position on guanine. The greater antitumor activity of the compounds in the modified nitrosourea class, when compared to the parent nitrosourea, has been attributed partly to their greater lipophilic character [43]. The latter subclass of compounds in this group are doxorubicin analogs, thought to inhibit DNA topoisomerase II and protein kinase C mediated signal transduction pathways [43]. The structural similarity of these latter compounds originates in their anthracene scaffold. The various congeners in this group do not appear to effectively affect growth inhibitory behavior, since they all exhibit similar activity patterns in the SVD space when compared to the complete set of 122 compounds. Three of the compounds within the group of anthracyclines share a dimethyl or diethyl amine group (308847, 142892 and 366140). Amonifide (308847) is a topoisomerase II inhibitor that acts as a DNA intercalator or binder [43], while pyrazoloacridine (366140) and hycanthone (142982) share an acridine moiety which may contribute to their similar activities.

The second group of compounds shares structural similarity with members of Group 1, but has SVD distance patterns different from the first group. Three of these compounds have aziridine or oxirane groups (6396, 9706 and 132313), four compounds are nitrogen mustards (762, 34462 and 344007) and one is a doxorubicin analog (269148). The diepoxides in the oxirane, dianhydrogalactitol (132313), are presumably responsible for its antitumor activity. Also within this group are two camptothecin analogs (643833 and 100880) and piperazinedione (135758), two of these compounds exhibit an alkylation capacity probably because of their chloride groups.

The third group (Group 3) includes sixteen compounds. This group includes two mitomycins (26980 and 56410), the only known natural compounds containing an aziridine ring [43]. These compounds alkylate guanine at the N-2 position in the DNA minor groove[43] and differ from one another only by a methyl group. With the exception of the topoisomerase II inhibitor 249992, the remaining compounds in this group are camptothecin analogs, that are thought to inhibit the DNA gyrase enzyme topoisomerase I. The strong structural similarity within the camptothecin derivatives is thus also exhibited in their SVD distance

patterns. Groups 4 and 5 consist of six and two camptothecin analogs, respectively. The cellular activities of the compounds in these two groups are sufficiently different from the larger set in Group 3 to include them as separate groups. The structural features responsible for this different activity are not clearly apparent. These compounds may exhibit similar activity patterns on the basis of solubility, or cell permeability.

Group 6 consists of only two compounds, the podophyllotoxin, Teniposide (122819) and the topoisomerase II inhibitor 301379. Although both of these compounds share structural similarity and activity patterns with the alkylating compounds in Group 1, their location adjacent to the group of Topoisomerase I agents suggests that their structural differences produce a distinctly different activity pattern.

Cluster 7 is a singlet, composed of aphidicolin glycinate (303812). This compound shares structural similarity with the camptothecin family, and its placement in a cluster near the camptothecin analogs in Groups 3, 4 and 5 suggests that its cellular activity may mimic that of Topoisomerase I inhibitors.

Twelve compounds are found in Group 8. Included in this set are the platinum containing, DNA intercalating, compounds tetraplatin (363812) and carboxyphthalatoplatinum (271674). These compounds contain a stabilizing cyclohexane group that may contribute to their distinctive activity patterns when compared to the three platinum containing compounds in Group 1. Seven nucleoside analogs appear within this Group (163501, 126771, 752, 71851, 71261, 118994 and 102816), most of which share a guanine or uracil moiety linked to a pentose. These compounds are thought to be directly incorporated into DNA [43]. The antibiotic acivicin (163501) and dichloroallyl-lawsone (126771) are thought to act as an inhibitor of pyrimidine biosynthesis, and their location within the family of nucleoside analogs is reasonable. The three doxorubicins that complete this group morpholinodoxorubicin (354646), cyanomorpholino-doxorubicin (357704) and N,N-dibenzyl duanomycin (268242), share a unique hexopyranosyl moiety. The two platinum containing alkylating agents and the three doxorubicin analogs act by directly damaging DNA, while the remaining compounds in this group are inhibitors of nucleotide synthesis, acting as DNA/RNA antimetabolites.

The antitubulin agents are found to cluster into five groups. The first group (Group 9) is composed of six antitubulin agents (330500, 332598, 153858, 49842, 609395, and 376128), one Topoisomerase II inhibitor (337766) and trityl cycteine (83265). The second group (Group 11) includes Taxol (125973) and a taxol derivative (608832). The third and fourth groups (Groups 12 and 13, respectively) include the cholchicines (757, 67574, 406042, 361792) and 33410. These compounds show weak pattern similarity to other anticancer agents, which suggests that these antitubulin agents share similar growth inhibitory mechanisms in the cell screen.

Group 10, which has an activity pattern that places it between the antitubulin Groups 9 and 11, consists of a nucleoside analog (19893), two amino acid analogs (153353 and 224131) and a folate analog (368390). Group 10 is the first cluster of compounds that lack close SVD distances to members of Groups 1-8. Thus its activity pattern lacks near SVD distances to groups containing alkylating agents and Topoisomerase I and II inhibitors with close SVD distances restricted mostly to members within its group. As will be shown later, this type of activity pattern may reflect agents that primarily act as inhibitors of nucleotide biosynthesis, rather than as DNA damaging agents.

An equally distinct activity pattern is also found for the antifolate compounds composing groups 14 and 15. Group 14 consists of Methotrexate (740) and the folate-analog (174121), while Group 15 includes the antimetabolites 633713 and 352122. It should be noted that in general, clustering of compounds in this subgroup is based largely on their SVD distance dissimilarities, rather than similarities, to the other members in the set of 122 compounds.

Groups 16-22 are all comprised of single compounds, all of which are nucleosides that act as antimetabolites of nucleotide biosynthesis. As with the folate analogs discussed above, their activity patterns are sufficiently unique for these compounds to share no pattern similarities with any of the standard 122 agents.

Folate analogs complete the final three groups. Groups 23 and 24 consist of single compounds (139105 and 623017, respectively), while Group 25 consists of three folates (184692, 134033 and 132483). These latter RNA/DNA antimetabolites have alcohols or ethers substituted at positions C-7 or C-11 of the parent compound

that may contribute to their increased water solubility and unique activity pattern.

The results described here are consistent with earlier classifications by Koutsoukos et al. [27] and van Osdol et al. [29] that divided these compounds into two large clusters. Our analysis finds a similar division of compounds, while providing further subclustering of compounds within these two major divisions. The largest division consists of compounds with the most similar activity patterns, compounds which appear at the top of Table 1, comprised primarily of DNA-damaging agents (groups 1-8). Compounds in the lower portion of Table 1 comprise the second major division and act by targeting a biosynthetic pathway or part of the mitotic machinery.

Each of the groups described above can be further examined for their average activity patterns across the 60 tumor cell lines. Figure 2 displays the mean activity for the 25 different groups across all 60 tumor cell lines. These results provide an indication of the diversity of activity patterns associated with the 25 clusters identified above, and can be used to identify which groups of compounds are more or less active against individual cell lines or within panels of cells. The results in Figure 2 are displayed according to the cluster order in Table 1, from Group 1 to Group 25. Thus alkylating agents (Group 1) appear as the first row and the activity pattern of Group 25 appears as the last row in this figure. Groups with positive mean activity patterns (greatest sensitivity) are displayed from least, to intermediate, to greatest, in orange, red and brown, respectively. Groups with negative mean activity patterns (least sensitivity) are shown, from least to intermediate, to greatest, in light blue, blue and dark blue, respectively. Groups with near zero mean activity patterns are shown in green.

Examination of the mean activity patterns for the 25 clusters obtained from the cubic clustering algorithm in SAS can be used to qualitatively assess differences between each group. The agents within Groups 1-3 exhibit a uniformly weak mean activity pattern across all 60 cell types. Groups 4 and 5 begin to exhibit a more diverse pattern, with a stronger sensitivity to the panel of CNS cells, as well as selected RENAL, LEUKEMIA and BREAST cells. Group 4 is composed of five camptothecin analogs that have an apparent, albeit weak, selectivity for the CNS panel of cells, with a strong activity against the single BREAST-ADR cell line. Group 6 is characterized by a strong insensitivity to the BREAST-ADR cell

line, while Group 7 exhibits a strong sensitivity to RENAL-ACHN, LEUKEMIA-HL60 and NLC-H460. Group 8 appears to have uniform activity against all cell lines. Groups 9-13, the antitubulin active agents, display a modest positive activity against all members of the BREAST panel, with the exception of BR-T47D, which displayed a strong insensitivity to these agents and to selected cells in the COLON panel. Groups 14-16 showed a mixed activity pattern within the BREAST and COLON panels, with both positive and negative mean activities within these cells. Group 17 displays a consistently positive activity against most of the cells within the BREAST panel and only the COLON-HCT15 cells. The single compounds in Groups 18-24, as well as the three compounds in Group 25, exhibited a widely diverse range of activity patterns. When compared to all the clusters identified in this analysis, Groups 18, 20 and 24 had the strongest positive activity patterns against COLON-HCC2998, MELONOMA-SK-MEL12 and NLC-EKVX, respectively. Cells with the least sensitivity to the 122 standard agents are: NLC- EKVX, BREAST-T47D, HS578T, and MDA231, OVARIAN-OVCAR4, RENAL- RXF393 and CNS-SNB75.

Our analysis can be used to cluster members of the 60 cell panel according to their response to the 122 standard anticancer agents. In contrast to the previous analysis where 122 agents were examined for their activity pattern across the 60 cell lines, a similar analysis can be performed whereby the 60 cell lines are examined for their activities against the 122 standard agents. Clustering of the cell types on this basis can be used to identify each cell type's differential response to these standard anticancer agents. Fifteen clusters are obtained using the cubic clustering analysis (CCC) within SAS. Figure 3 displays a cladogram for clusters obtained in this analysis, with each branch labeled and color coded according to cell type. Cells are initially separated into two major branches, with one branch consisting of 15 cell types, while the remaining 45 cell types appear in the other major branch.

The smaller of the two major branches appears at the rightmost portion of Figure 3, and is subdivided into four clusters. The largest of these four clusters consist of RENAL cell types, with UO-31, 786-0, ACHN, CAKI-1 and RXF-393 along with two MELANOMA cell lines, LOX-IMVI and M14. Four of the five RENAL cells in this panel are know to exhibit multidrug resistance (MDR). MDR is a known complication of cancer therapy associated with either an increased expression of the P-170 membrane glycoprotein MDR1 or the

presence of the multidrug resistance protein (MRP) [64, 65]. Both of these mechanisms act by lowering the effective drug concentration, enhancing drug efflux [43] and reducing drug efficacy. The remaining three sub-branches within this major branch are comprised of four LEUKEMIA, two NLC, one CNS and one MELANOMA cell type. The LEUKEMIA cell line has the greatest average sensitivity in mean deviation ($\Delta x = [\log GI_{50}] - \langle \log GI_{50} \rangle$) for the 122 standard agents. The LEUKEMIA cell type SR appears as a singlet, thus having no comparable cell type with a similar response to the 122 standard agents.

The larger of the two major branches found in this analysis is clustered into 4 sub-branches, which are further divided into 17 branches. The leftmost sub-branch (as viewed in Figure 3) is divided into 7 clusters. The largest cluster in this group consists of seven cell types, appearing as the left-most branch of the cladogram. This cluster includes three OVARIAN, two NLC and one MELANOMA cell type. Adjacent to this cluster are four branches comprised of only a single cell type: (RE)SN12C, (CNS)SF-268, (BR)BT-549 and (ME)MALME-3M. Two BREAST cell types (T-47D and MCF7) along with the LEUKEMIA cell line RPMI-8226 appear in the next cluster. Membership in this leftmost sub-branch is completed by a cluster comprised of only two OVARIAN cell types (SK-OV-3 and OVCAR-8) and the singlet (NLC)HOP-92. The remaining clusters in this major sub-branch consist primarily of NLC, COLON, BREAST and MELANOMA cell types. Within the clusters formed by these cell types, a clear separation according to these panels is not apparent based on their response to the 122 standard agents. An apparent coherence between the COLON, BREAST and LEUKEMIA panels is clearly indicated, however the basis for this clustering is not evident. These results indicate that
many tumor cell types, both within and between different panels, exhibit similar sensitivities to the set of 122 compounds studied here. Additional studies with a larger set of test compounds will be needed to more thoroughly determine which cell types share the most similar response patterns.

Prediction of MOAs

Mechanism of action classifications can be based on applications of a wide range of statistical tools [35, 36, 38]. The results in Table 1 show that there is a substantial similarity between the clusters of compounds based on GI_{50} activity patterns and their classification based on their previously assigned MOAs. Yet, subclusters interspersed between clusters of a given MOA are observable, which call for a more systematic analysis of the degree of correlation between the GI_{50} data and MOAs. To this aim we performed the following analysis: Mean activity fluctuation vectors in the SVD space were found for each of the six MOAs using

$$< \mathbf{a}^* >_{MOA} = \sum_i \mathbf{a}_i^* / N_{MOA} \qquad (5)$$

Here N_{MOA} is the number of agents exhibiting a given MOA, and the summation is performed over this particular subset of agents. The average activity patterns are thus obtained for each MOA. The departure of the behavior $\mathbf{a}_i^*$ of individual agents from these averages are examined for an assessment of the accuracy of the MOAs assigned to the different agents. The deviation of each drug from the mean activity fluctuation vector for the six MOA classes is thus

$$\Delta \mathbf{a}_i^*{}_{MOA} = \mathbf{a}_i^* - < \mathbf{a}^* >_{MOA} \qquad (6)$$

The smallest of the six distances obtained for each drug is used to identify its most likely MOA. Application of this test to all compounds in the training set of 122 standard agents shows that the correct MOAs are assigned with an average accuracy level of 96.7%. Column 2 in Table 2 summarizes the results for the six different MAOs. Weinstein et al. [26] obtained an accuracy level of 91.5% by using neural network model, and 85.8% by linear discriminant analysis [26].

The accuracy of the MOA assignments for anticancer agents has additionally been examined by jackknife tests. The jackknife test, also called the leave-one-out test [45], is a method often utilized for small samples which cannot be divided into training and testing sets without loss of information. In this procedure each compound to be tested is removed from the training dataset and the identification of the activity fluctuation $\Delta \mathbf{a}_i^*{}_{MOA}$ for each MOA is carried out using the GI_{50} data of the remaining 121 drugs. The most probable

MOA of the test compound is then predicted using the same distance criteria (eq 6), with the basic difference that the mean fluctuation vectors $\langle \mathbf{a}^* \rangle_{MOA}$ are now extracted from a set of data excluding the test compound. The average accuracy level reached by this method was 84.4%. A summary of these results is presented in the third column of Table 2. The mispredicted compounds and their predicted MOA's are listed in Table 3. Most of the 19 mispredicted compounds were classified as topoisomerase II agents or DNA/RNA antimetabolites, with the majority of these agents predicted to behave as alkylators.

Discussion

NCI's 60 cell line screening assay provides a measure of growth inhibition for human cancer cells exposed to candidate anticancer compounds. Activity data accumulated in these screens can be used to group agents that exhibit similar activity patterns across a broad variety of tumor cell lines. Compounds grouped according to pattern similarities can be further examined for possible relationships between their activities, their chemical substructures and/or MOAs. The results presented here apply the standard statistical method of singular value decomposition (SVD) to the $\log(GI_{50})$ data to define measures of distances between compounds in a space that best distinguishes their similarities and dissimilarities. Hierarchical clustering of these SVD-derived distances divides these 122 compounds into 25 groups. The first eight groups are predominantly formed by DNA-damaging agents, while the latter seventeen groups (9-25), mostly consist of agents that inhibit nucleic acid biosynthesis or mitosis. Compounds in the first class comprise MOAs assigned as alkylators, and inhibitors of topoisomerases I and II, along with a few DNA antimetabolites, while the latter class is dominated by antimitotic agents and antimetabolites.

DNA damaging agents (Groups 1-8), when observed together, exhibit strongly similar activity patterns. Agents such as DNA alkylators and DNA metalators (platinum agents) are equally effective against slowly dividing or non-dividing cells (termed G_0 cells). Since strong pattern similarities are observed among alkylators and platinum analogs, it is reasonable to conclude that these compounds have comparable activities against all cell types, as evidenced by the uniform activity pattern for these groups. Thus compounds

that act directly on DNA, either by cross-linking or less directly by inhibiting enzymes responsible for processing DNA (i.e. unwinding) fall into this first group. While alkylating agents would be expected to be included in the class of DNA-damaging agents, the present finding that topoisomerase inhibitors behave similarly to alkylating agents is unexpected. However, inhibition of topoisomerases result in DNA damage, with repair modulated by the impact of the damage. Earlier studies have found that some topoisomerases are constitutively expressed at relatively constant levels throughout the cell cycle, even in cells that are not actively dividing [46]. Thus inhibitors of topoisomerases may potentially be active in tumors that have low growth fractions [43] and as a result exhibit cytotoxic behavior similar to alkylating agents.

The second major class of compounds identified in our analysis acts against the enzymatic machinery required for cell division. Most of these compounds inhibit purine or pyrimidine biosynthesis or act as antitubulin agents. Evidence to support this claim can be found in the crystallographic complexes between biosynthetic enzymes and ligands that are either identical to those included in the set of 122 compounds, or close structural analogs. Although it is not our intention here to present a systematic analysis of structural data in support of this claim, Appendix A summarizes our survey of the crystallographic database of proteins complexed with ligands that bear strong structural similarity to many of the antimetabolite agents in the set of 122 compounds.

A strong correspondence was not observed between specific MOAs of compounds assigned to each cluster. For example, alkylating agents and topoisomerase I and II inhibitors appear in most of the first 8 clusters. The results of this analysis are, however, sufficiently meaningful to yield a MOA prediction accuracy greater than 84%. Inspection of the subclusters obtained from this analysis finds compounds that both share and lack structural similarity.

Many approaches are available for classification of compounds by chemical structure [30, 52]. Some approaches are based on one-dimensional (1-D) global features such as polarizability, molecular weight, and number of hydrogen bond donors/acceptors [53, 54]. Alternative approaches attempt to maximize a

selection of 2-dimensional (2-D) and 3-dimensional (3-D) indices assigned to each compound [55-57]. Some of the more commonly used descriptors are based on chemical formula [57], 2-D topological similarity [58-61] and 3-D superposition [62]. Using sets of indices representative of these descriptors, compounds can be assigned a 'fingerprint' which can be used for assessing similarities within groups of compounds [17]. Clusters of the 122 compounds examined here, based on a set of 54 1-D descriptors available in the Cerius package and based on 2-D SMILES descriptors, found no statistically significant correlation with the activity patterns from the screening assay. Taken separately or together, no combination of these 1-D or 2-D descriptors could be found to produce a statistically significant correlation with the activity patterns observed for the 122 agents examined here. Although examination of Figure 1 provides clear evidence that many compounds within each group have common substructural features, a systematic means of assigning these compounds to these groups, on the basis of 1-D and 2-D descriptors alone, was not apparent. These results are consistent with widespread observations such as those of Brown et al. [58], where small chemical modifications can result in quite different biological responses. The family of camptothecins offers a clear example of such behavior, i.e. small differences in the parent structure resulted in quite different activity patterns. Our results emphasize the importance of assessing structural information together with screening data to assess biological activity.

One important question arises about studies such as that presented here - What is the effect of data errors on the results? Single compounds, such as those clustered in Groups 16-24 above, are easily distinguished in this type of analysis. Hierarchical clustering of SVD distances alone identifies these singlets on the basis of their position in a separate branch of the tree. The additional classification based on pairwise differences in SVD distances with respect to the whole set of compounds can be further used to determine whether compounds isolated in a single branch of the tree have an important different activity pattern, or lack any such feature.

Measurement errors that appear in the reported $\log(GI_{50})$ values represent another type of error. These errors result from experimental conditions as well as errors in data reporting. In an attempt to address the importance of these types of errors on our results, the current dataset was perturbed with

random noise and the SVD distances were recalculated. Figure 4 displays the results of perturbing the current set of $\log(GI_{50})$ values by an error that ranges from zero to 40%. The ordinate in Figure 4 represents the correlation coefficient [63] between the matrix of SVD distances calculated for the unperturbed and perturbed datasets. There we see that perturbing the existing data with 20% error yields an SVD distance matrix whose entries are still correlated with the original data with a correlation coefficient of 0.9. By contrast, a 40% error produces a correlation coefficient near 0.7. From this analysis we believe that data error in the range of 10 to 20% should yield results extremely similar to those reported here. The actual error in this data is difficult to establish. An estimate of the maximum error can be obtained by calculating the coefficient of variation (C.V. = $\sigma/\overline{\log(GI_{50})}$) for the $\log(GI_{50})$ values obtained for each compound. The variance (σ) is estimated therein as the squared sum of Δx_{ij} calculated in equation 3. This method yields a coefficient of variation of 0.87 (or a percentage error of 13%), which according to Figure 4, corresponds to a correlation coefficient of 0.95. We conclude that the results of our analysis are robust enough to sustain errors lower than 15% without significant degradation. The experimental data used in our study include results from multiple replicate analysis performed between 2 to 50 replicates, which would reduce the measurement noise.

Based on the above observation that selected cell types could be clustered according to their response to the 122 standard agents, we explored whether differences in SVD distance clusters would occur from analyses based on subsets of selected cell types that are known to exhibit multidrug resistance (MDR). Based on the relative expression of MDR-1 mRNA and the immunocytochemical characterization of P-glycoprotein expression [66], eight MDR1 expressing cell types are identified: HCT-15(CO), SF-295(CNS), HOP-62(NLC), UO-31(RE), A498(RE), ACHN(RE), CAKI-1(RE) and RXF-393(RE). This selection conforms most closely to those cells exhibiting the highest rhodamine efflux measurements as posted on the Developmental Therapeutics' web page (http://dtp.nci.nih.gov). Clustering analysis was performed using (i) the $\log(GI_{50})$ values from the 8 MDR1 expressing cell lines, and (ii) the $\log(GI_{50})$ values from the 52 non-MDR1 expressing cell lines in the screen. The analysis based on the 52 non-MDR1 expressing cells clustered compounds qualitatively similar to that obtained for the complete set of 60 cell lines. The analysis performed on the 8 MDR1 expressing cells

found that the activity patterns within this group had similar SVD distances, and their activity pattern with respect to their response to the 122 standard agents was quite similar to that found for the previously classified DNA-damaging agents. In particular, the antitubulin agents found in Groups 9, 11, 12 and 13 exhibit SVD distances that are similar to the members of the DNA damaging agents in Groups 1-8. In addition to this subset of antimitotic agents, the antimetabolites found in Groups 14-25 also display SVD distance patterns that reflect patterns closely resembling that of the DNA damaging agents. This result is consistent with the view that MDR is associated with the increased efflux of etoposides, anthracyclines (topoisomerase II inhibitors), colchicines and vinca alkaloids (antimitotic agents) [43, 44], and also demonstrates that agents that inhibit nucleotide biosynthesis are also affected. The result of multidrug resistance is a more uniform activity pattern across all cell panels, a feature characteristic of DNA damaging agents.

The results presented herein can be contrasted with those available from the web-accessible program COMPARE. Comparisons of the Pearson Correlation Coefficients (PCC) of the 122 standard agents with the SVD distances were not statistically significant. The strongest differences were observed for compounds with statistically significant PCC values (above PCC=0.38, $p<0.05$, $n=59$) that were found to have large SVD distances. For example, the highest PCCs for the set of 122 standard anticancer agents are found for most of the compounds classified, by our analysis, as DNA damaging agents. A COMPARE analysis based on a 'seed' selected from compounds in Groups 1-6 found statistically significant 'hits' for over half of the 122 standard agents, many of which were found to have large SVD distances. Instances where statistically significant PCC values corresponded to near SVD distances were observed for compounds in Groups 8, 10, 11 and 12 and the single compounds in Groups 14-24. Thus better agreement between the two approaches was found for compounds that inhibit nucleic acid biosynthesis or mitosis. While it is not our intention here to produce a detailed comparison of these two methods, it is clear that both approaches yield varying degrees of agreement, depending on the compound of interest.

In summary, statistical clustering tools have been used to analyze the growth inhibitory potency data available from the NCI's 60 tumor cell line screen. Analysis of the results for 122 standard anticancer agents finds that this set of

compounds can be clustered according to screening patterns into 25 groups, with eight of these groups consisting of DNA damaging agents and the remaining groups consisting of agents that act either to inhibit nucleotide biosynthesis or mitosis. Structural similarities are found between compounds assigned to these two broad categories. Clustering of the cell types based on their response to the 122 standard agents divided the cells into two major branches which were further sub-divided into 21 groups. Strongest within-panel responses were found for the RENAL, OVARIAN and LEUKEMIA panels. The current analysis provides a reference for evaluating larger data sets of compounds for similarities in their screening patterns with respect to the standard 122 anticancer agents. Analyses of these larger data sets may be able to relate more precisely chemical substructure to activity.

Acknowledgements

The authors are grateful for discussions with Drs. Anne Monks and Dominic Scuderio about the cell screening data. We would also like to acknowledge the TUBITAK fellowship of O. Keskin, the IRSP program at SAIC Frederick and Grant DAMD17-98-1-8323 from the U.S. Army.

Table 1. Compounds Ordered According to Pattern Similarity.

Cluster	Name	NSC	MOA	structural group
1	teroxirone	296934	1	epoxide
1	AZQ	182986	1	aziridine
1	CHIP	256927	1	platinum
1	cis-platinum	119875	1	platinum
1	carboplatin	241240	1	platinum
1	hepsulfam	329680	1	alkane sulfonate
1	Yoshi-864	102627	1	alkane sulfonate
1	Busulfan	750	1	alkane sulfonate
1	cyclodisone	348948	1	alkane sulfonate
1	clomesone	338947	1	alkane sulfonate
1	guanazole	1895	6	
1	pyrazoloimidazole	51143	6	
1	ftorafur (pro-drug)	148958	5	
1	hydroxyurea	32065	6	hydroxyurea
1	melphalan	8806	1	nitrogen mustard
1	chlorambucil	3088	1	nitrogen mustard
1	br-propionyl piperazine	25154	1	nitrogen mustard
1	fluorodopan	73754	1	nitrogen mustard
1	mitozolamide	353451	1	nitrogen mustard
1	BCNU (Carmustine)	409962	1	nitrosourea-nitrogen mustard
1	spirohydantoin mustard	172112	1	nitrogen mustard
1	methyl CCNU	95441	1	nitrosourea-nitrogen mustard
1	chlorozotocin	178248	1	nitrosourea-nitrogen mustard
1	PCNU	95466	1	nitrosourea-nitrogen mustard
1	CCNU	79037	1	nitrosourea-nitrogen mustard
1	3-HP	95678	6	hydrazinecarbonthioamide
1	5-HP	107392	6	hydrazinecarbonthioamide
1	asaley	167780	1	nitrogen mustard
1	amonafide	308847	4	-
1	hycanthone	142982	1	-
1	pyrazoloacridine (PZA)	366140	4	acridine
1	oxanthrazole	349174	4	anthracene
1	anthrapyrazole derivative	355644	4	anthracene
1	rubidazone	164011	4	anthracene dione
1	doxorubicin (Adriamycin)	123127	4	anthracene-daunorubicin
1	daunorubicin	82151	4	anthracene-daunorubicin
1	deoxydoxorubicin	267469	4	anthracene-daunorubicin
1	VP-16	141540	4	podophyllotoxin
2	thio-tepa	6396	1	aziridine
2	triethylenemelamine	9706	1	aziridine
2	dianhydrogalactitol	132313	1	epoxide
2	nitrogen mustard	762	1	nitrogen mustard
2	uracil nitrogen mustard	34462	1	nitrogen mustard
2	piperazine analog	344007	1	nitrogen mustard
2	piperazinedione	135758	1	piperazine

2	camptothecin derivative	643833	3	camptothecin
2	camptothecin, Na salt	100880	3	camptothecin
2	menogaril	269148	4	anthracene-daunorubicin
3	mitomycin C	26980	1	mitomycin
3	porfiromycin	56410	1	mytomycin
3	camptothecin	94600	3	camptothecin
3	camptothecin derivative	95382	3	camptothecin
3	camptothecin derivative	107124	3	camptothecin
3	m-AMSA (Amsacrine)	249992	4	anthracene
3	camptothecin derivative	295501	3	camptothecin
3	camptothecin derivative	606173	3	camptothecin
3	camptothecin derivative	364830	3	camptothecin
3	camptothecin derivative	374028	3	camptothecin
3	aminocamptothecin	603071	3	camptothecin
3	camptothecin derivative	606172	3	camptothecin
3	camptothecin derivative	606985	3	camptothecin
3	camptothecin derivative	610457	3	camptothecin
3	camptothecin derivative	610458	3	camptothecin
3	camptothecin derivative	618939	3	camptothecin
4	camptothecin derivative	249910	3	camptothecin
4	camptothecin derivative	606947	3	camptothecin
4	camptothecin derivative	606499	3	camptothecin
4	camptothecin derivative	610456	3	camptothecin
4	camptothecin derivative	610459	3	camptothecin
4	camptothecin derivative	629971	3	camptothecin
5	camptothecin derivative	176323	3	camptothecin
5	camptothecin derivative	295500	3	camptothecin
6	VM-26 (Teniposide)	122819	4	podophyllotoxin
6	mitoxantrone	301739	4	anthracene
7	aphidicolin glycinate	303812	6	aphidicolin
8	tetraplatin	363812	1	platinum
8	carboxyphthalatoplatinun	271674	1	platinum
8	acivicin	163501	5	amino acid analog
8	dichlorallyl lawsone	126771	5	napthoquinone
8	thioguanine	752	6	guanine
8	alpha-TGDR	71851	6	guanine
8	beta-TGDR	71261	6	guanine
8	inosine glycodialdehyde	118994	6	guanine
8	5-azacytidine	102816	5	cytidine
8	cyanomorpholinodoxorubic	357704	1	anthracene-daunorubicin
8	morpholinodoxorubicin	354646	3	anthracene-daunorubicin
8	N, N-dibenzyl daunomycin	268242	4	anthracene-daunorubicin
9	macbecin II	330500	6	lactone
9	rhizoxin	332598	2	macrolide

9	maytansine	153858	2	macrolactam
9	vinblastine sulfate	49842	2	vinca alkaloid
9	halichondrin B	609395	2	polyether macrolide
9	trityl cysteine	83265	2	tri-phenyl
9	bisantrene HCL	337766	4	anthracene
9	dolastatin 10	376128	2	modified peptide
10	L-alanosine	153353	5	aspartate analog
10	N-(phosphonoacetyl)-L-aspartate	224131	5	aspartate analog
10	5-fluorouracil	19893	5	uracil analog
10	brequinar	368390	5	folate analog
11	taxol	125973	2	taxane
11	taxol derivative	608832	2	taxane
12	colchicine derivative	33410	2	colchicine
12	allocolchicine	406042	2	colchicine
12	thiocolchicine	361792	2	colchicine
13	colchicine	757	2	colchicine
13	vincristine sulfate	67574	2	vinca alkaloid
14	methotrexate	740	5	folate analog
14	methotrexate derivative	174121	5	folate analog
15	L-ornithine	633713	5	folate analog
15	trimetrexate	352122	5	folate analog
16	thiopurine	755	6	purine
17	5-aza-2'-deoxycytidine	127716	6	cytidine
18	2'-deoxy-5-fluorouridine	27640	6	uridine
19	ara-C	63878	6	uridine
20	5,6-dihydro-5-azacytidine	264880	5	cytidine
21	pyrazofurin	143095	5	pyrazofurin
22	cyclocytidine	145668	6	cytidine
23	Baker's antifol soluble	139105	5	folate
24	an antifol	623017	5	folate analog
25	aminopterin derivative	184692	5	folate analog
25	aminopterin derivative	134033	5	folate analog
25	aminopterin derivative	132483	5	folate analog

Table 2. Performance of SVD Analysis for Determining MOAs

Mechanism of action*	Success % training set	Success % Prediction set
1 (35)	97	97
2 (13)	92	85
3 (24)	96	96
4 (15)	100	87
5 (19)	100	63
6 (16)	94	63
AVG (122)	96.7	84.4

*MOA

1 – Alkylating

2 – Antimitotic

3 – Topoisomerase I inhibitors

4 – Topoisomerase II inhibitors

5 – RNA/DNA Antimetabolites

6 – DNA Antimetabolites

Table 3. MOA Classification for Incorrectly Predicted MOAs

NSC Number	Name	Assigned MOA	Predicted MOA
357704	cyanomorpholinodoxorubicin	1	3
153858	maytansine	2	6
67574	vincristine sulfate	2	6
354646	morpholinodoxorubicin	3	4
268242	N,N-dibenzyl daunomycin	4	1
366140	pyrazoloacridine	4	1
148958	Ftorafur	5	6
102816	5-azacytidine	5	4
264880	5,6-dihydro-5-azacytidine	5	1
174121	methotrexate derivative	5	6
139105	Baker's soluble antifol	5	2
132483	aminopterin derivative	5	3
623017	an antifol	5	6
63878	ara-C	6	1
27640	2'-deoxy-5-fluorouridine	6	1
127716	5-aza-2'-deoxycytidine	6	4
330500	Macbecin II	6	1
95678	3-HP	6	1
32065	hydroxyurea	6	1

Figure Captions

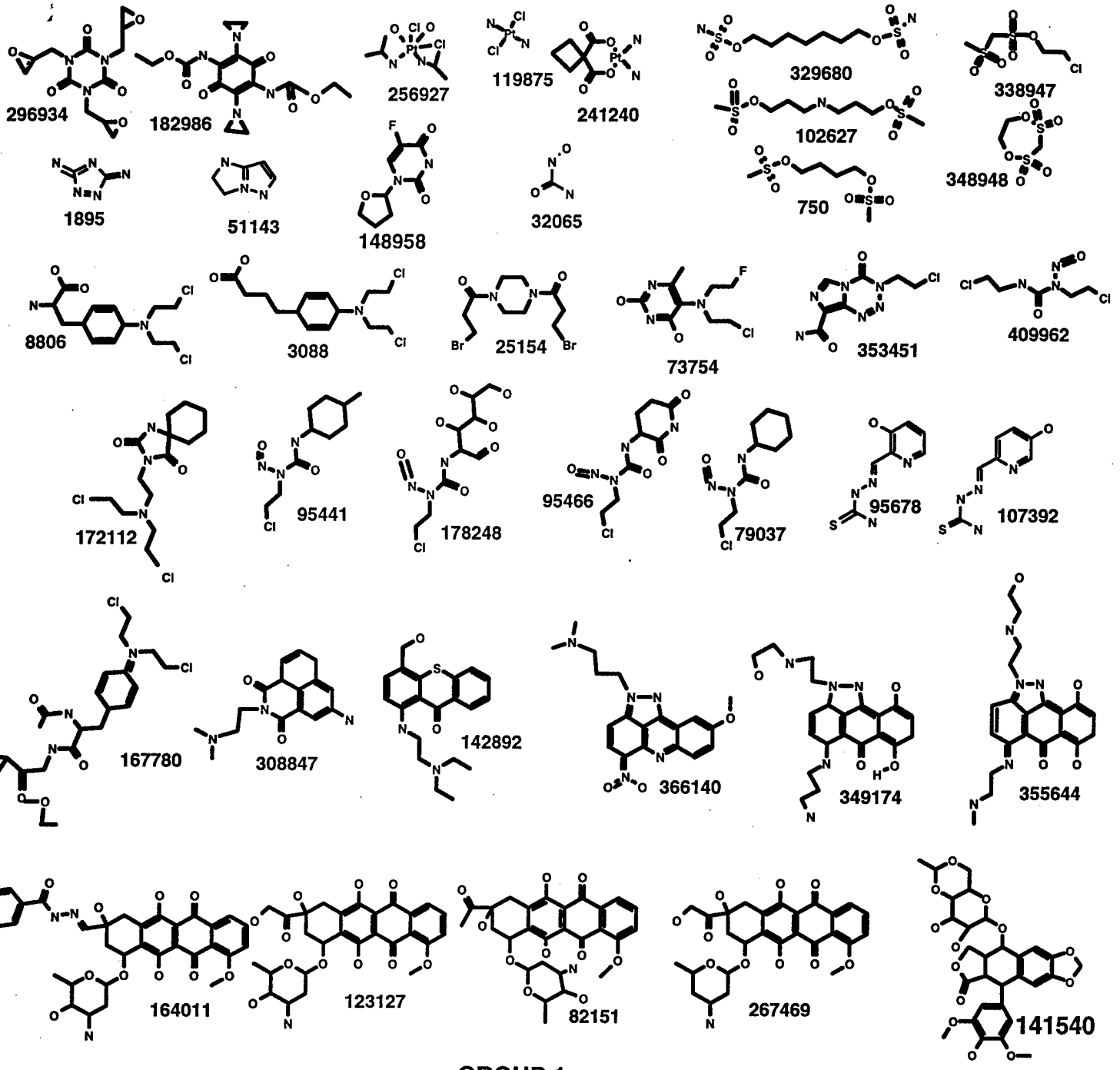
Figure 1. Two-dimensional representations of the chemical structures of the 122 compounds analyzed in this study. Compounds are ordered into 25 groups as described in the text. Structurally similar compounds are displayed together within each group. This figure has been prepared using the ISIS/DRAW software package.

Figure 2. Average activity across the 60 cell lines for compounds in each of the 25 groups. Panels of cells are ordered from bottom to top as follows: CNS, PROSTATE, MELANOMA RENAL, LEUKEMIA, OVARIAN, BREAST, COLON and NLC. Groups with a positive mean activity pattern are displayed from least, to intermediate, to greatest, in orange, red and brown, respectively. Groups with negative mean activity patterns are shown, from least to greatest, in light blue, blue and dark blue, respectively. Groups with mean activity patterns near zero are shown in green.

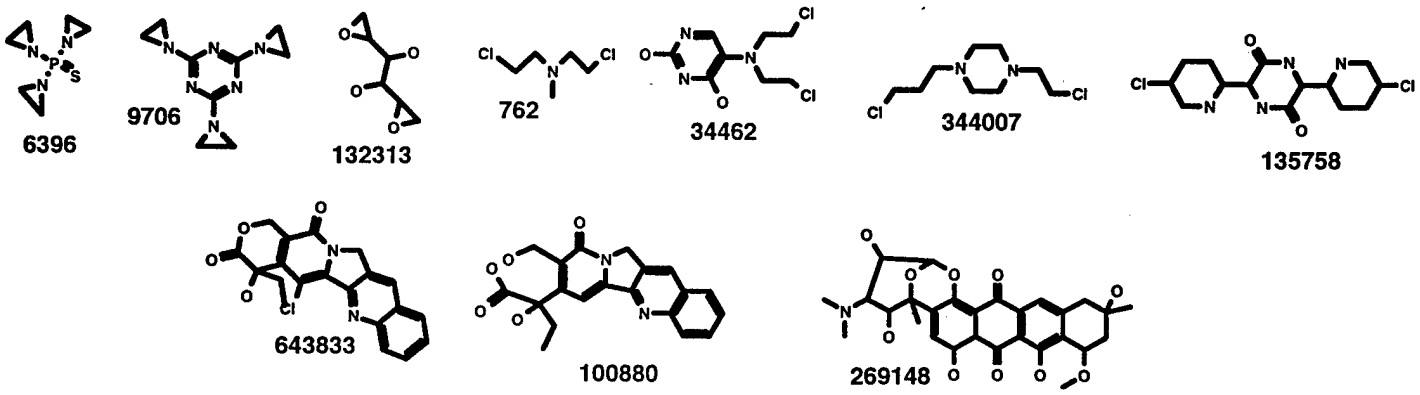
Figure 3. Cladogram of SVD distances for the 60 cell types determined from the activity data for the standard 122 anticancer agents. Branch labels are colored according to cell panels; black:non-small cell lung carcinoma(NLC), light green:COLON, magenta:LEUKEMIA, red:OVARIAN, dark green:RENAL, brown:MELANOMA, light blue:PROSTATE, black:CNS. (Note that the color black has been used for both NLC and CNS.) The abbreviations for each panel also appear in the label for each branch. The GROWTREE Utility from the GCG software package has been used to generate this figure. Cluster assignments, from left to right, are as follows; Cluster 1: (ME)UACC-62, (OV)OVCAR-5, (OV)OVCAR-4, (NLC)NCI-H322M, (OV)IVGROV1, (NLC)A549/ATCC, (RE)SN12C. Cluster 2: (CNS)SF-268. Cluster 3: (BR)BT-549. Cluster 4: (ME)MALME-3M. Cluster 5: (BR)T-47D, (BR)MCF7, (LE)RPMI-8226. Cluster 6: (OV)SK-OV-3, (OV)OVCAR-8. Cluster 7: (NLC)HOP-92. Cluster 8: (ME)SK-MEL-5, (NLC)EKVX, (RE)TK-10, (CNS)SNB-19, (CO)SW-620, (LE)K-562. Cluster 9: (CO)HCT-15. Cluster 10: (PR)PC-3, (BR)MDA-231. Cluster 11: (BR)HS-578T, (CO)HT29. Cluster 12: (BR)MDA-N, (BR)MDA-435, (OV)OVCAR-3, (CO)COLO-205, (CO)HCC-2998, (CNS)SF-295. Cluster 13: (PR)DU-145. Cluster 14: (NLC)NCI-H226, (NLC)NCI-H23, (RE)A498, (ME)SK-MEL-28, (NLC)NCI-H460, (CNS)U251. Cluster 15: (CNS)SNB-75. Cluster 16: (ME)SK-MEL-2, (CO)KM12, (CO)HCT-116. Cluster 17: (BR)NCI/ADR. Cluster 18: (ME)M14, (RE)RXF-393, (MEL)LOX-IMVI, (RE)CAKI-1, (RE)ACHN,

(RE)786-0, (RE)UO-31. Cluster 19; (LE)MOLT-4, (LE)CCRF-CEM, (ME)UACC-257, (NLC)HOP-62. Cluster 20: (LE)SR. Cluster 21: (CNS)SF-539, (LE)HL-60(TB), (NLC)NCI-H522.

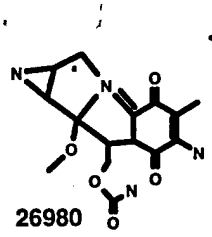
Figure 4. Sensitivity analysis of present SVD results. Correlation coefficients between the results found from SVD derived distances based on original $\log(GI_{50})$ data, and those based on the randomly perturbed $\log(GI_{50})$ data. The ordinate represents the percentage error introduced upon perturbation of the original data set.

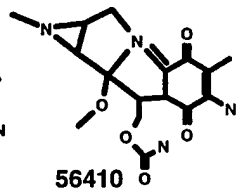


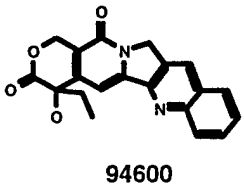
GROUP 1

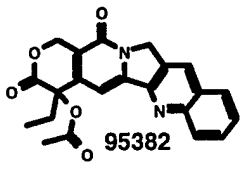


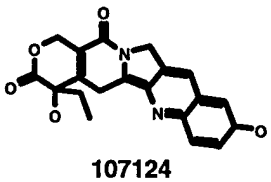
GROUP 2

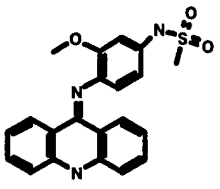

26980

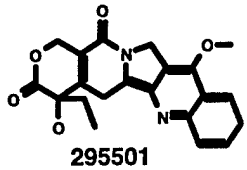

56410

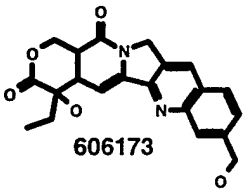

94600

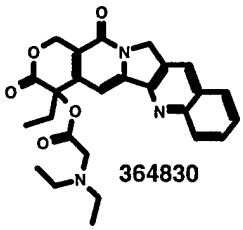

95382

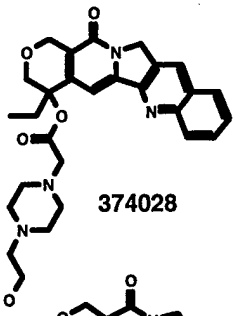

107124

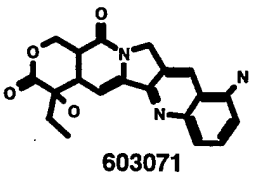

249992

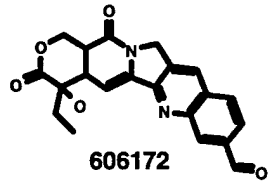

295501

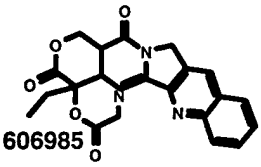

606173

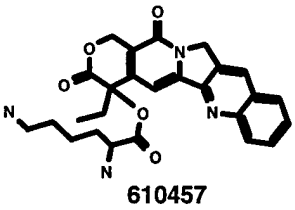

364830

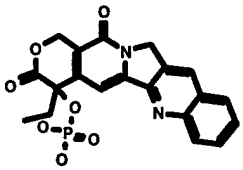

374028

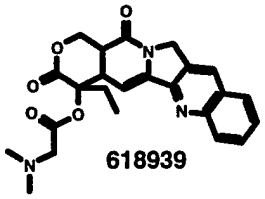

603071


606172

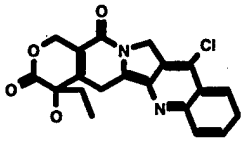

606985

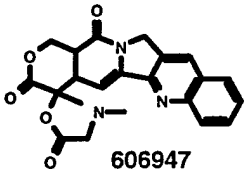

610457

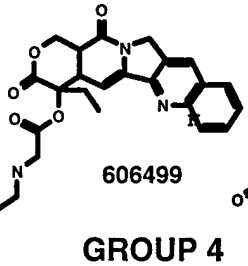

610458

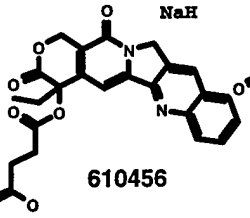

618939

GROUP 3

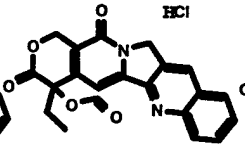

249910


606947

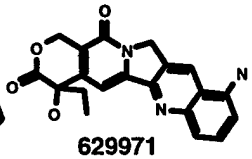

606499



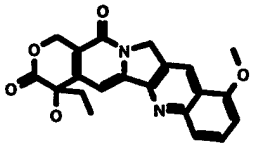
610456

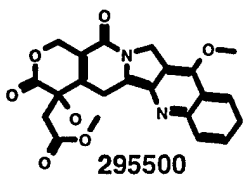


610459

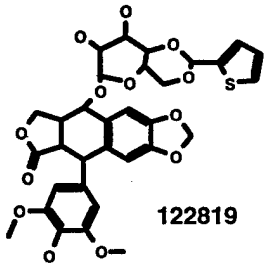

629971

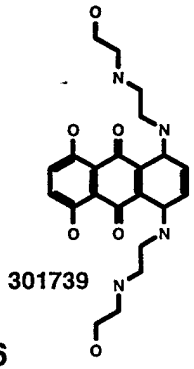
GROUP 4


176323

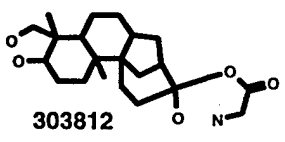

295500

GROUP 5

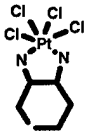

122819

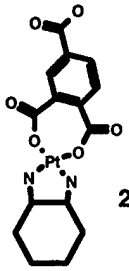

301739

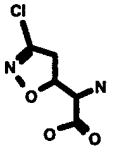
GROUP 6

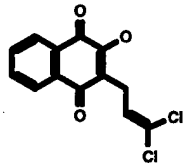

303812

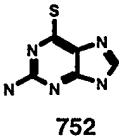
GROUP 7

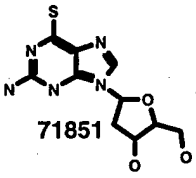

363812

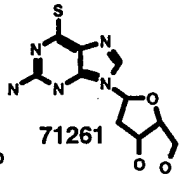

271674

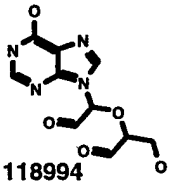

163501

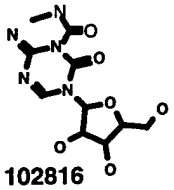

126771

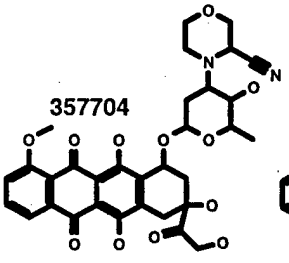

752

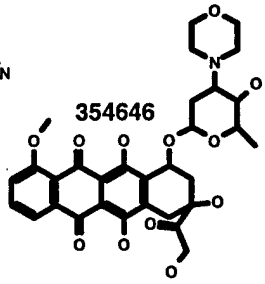

71851

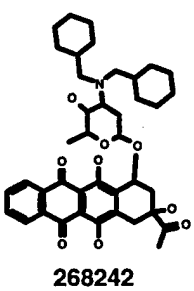

71261


118994

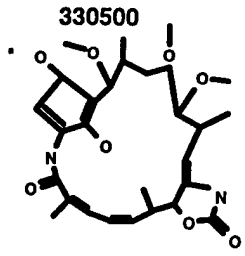

102816

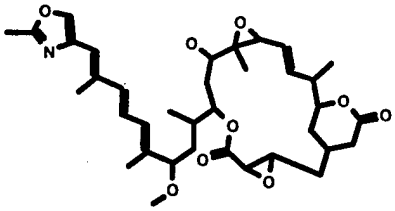

357704

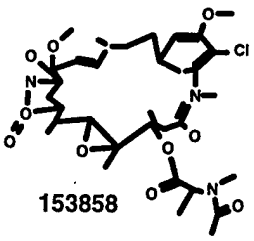

354646

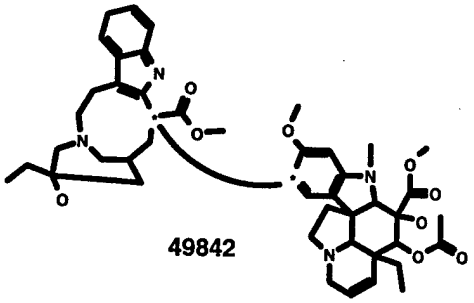

268242

GROUP 8

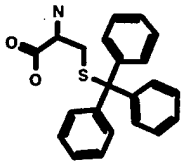

330500


332598

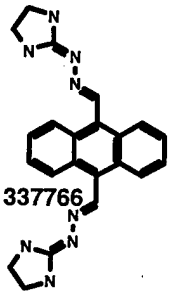

153858

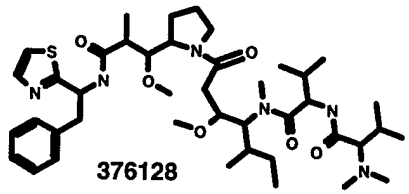

49842

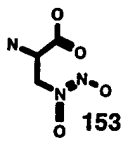
609395
too complex

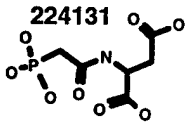

83265

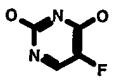
GROUP 9

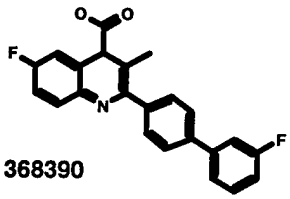

337766


376128

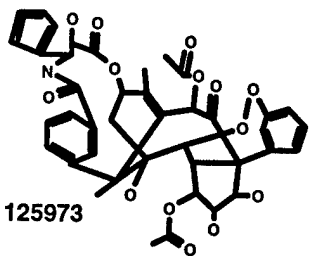

153353

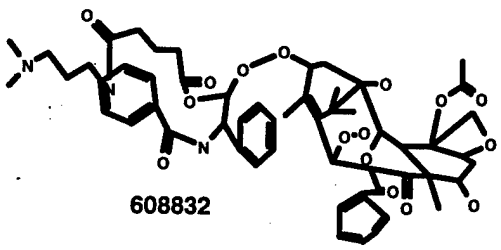

224131


19893


368390

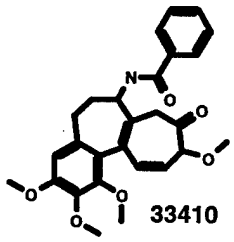
GROUP 10

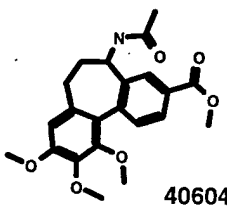

125973


608832

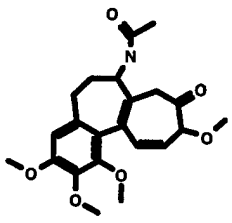
GROUP 11

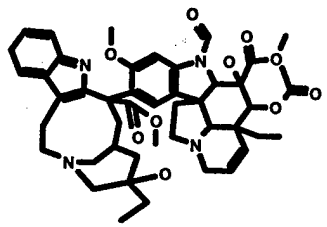
361792
too complex


33410

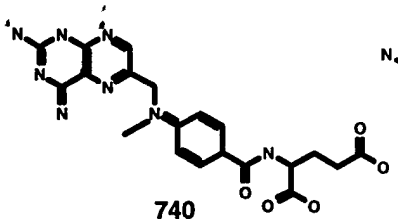

406042

GROUP 12

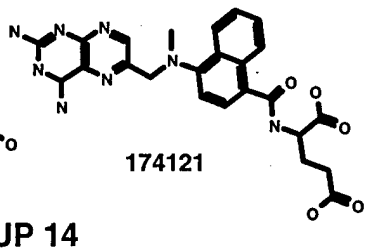

757


67574

GROUP 13

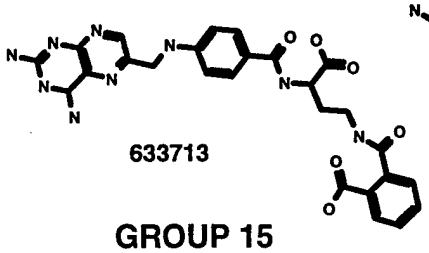


740

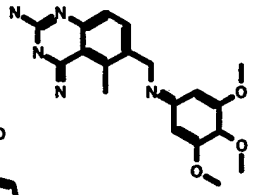


174121

GROUP 14

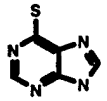


633713



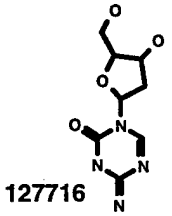
352122

GROUP 15



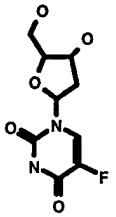
755

GROUP 16



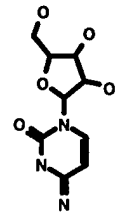
127716

GROUP 17



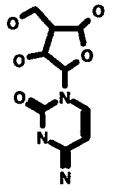
27640

GROUP 18



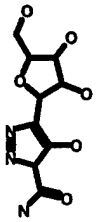
63878

GROUP 19



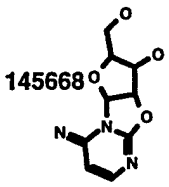
264880

GROUP 20



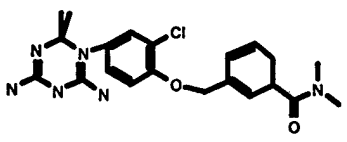
143095

GROUP 21



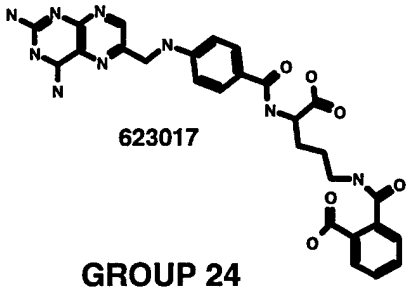
145668

GROUP 22



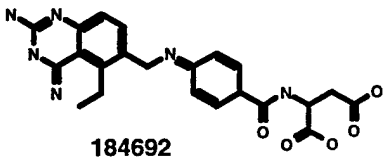
139105

GROUP 23

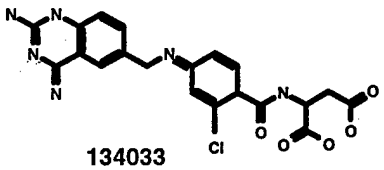


623017

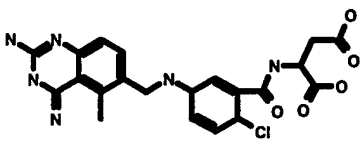
GROUP 24



184692

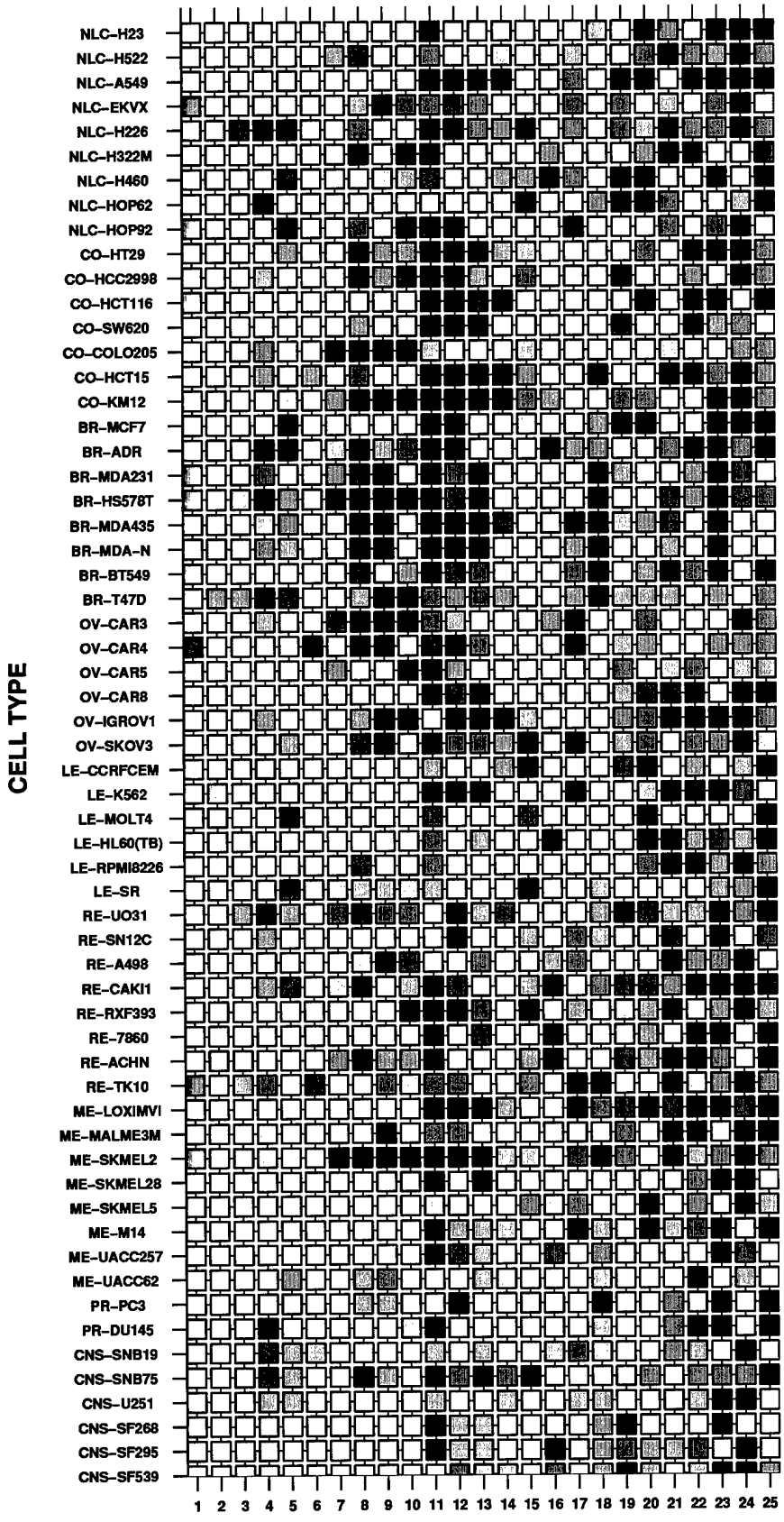


134033



132483

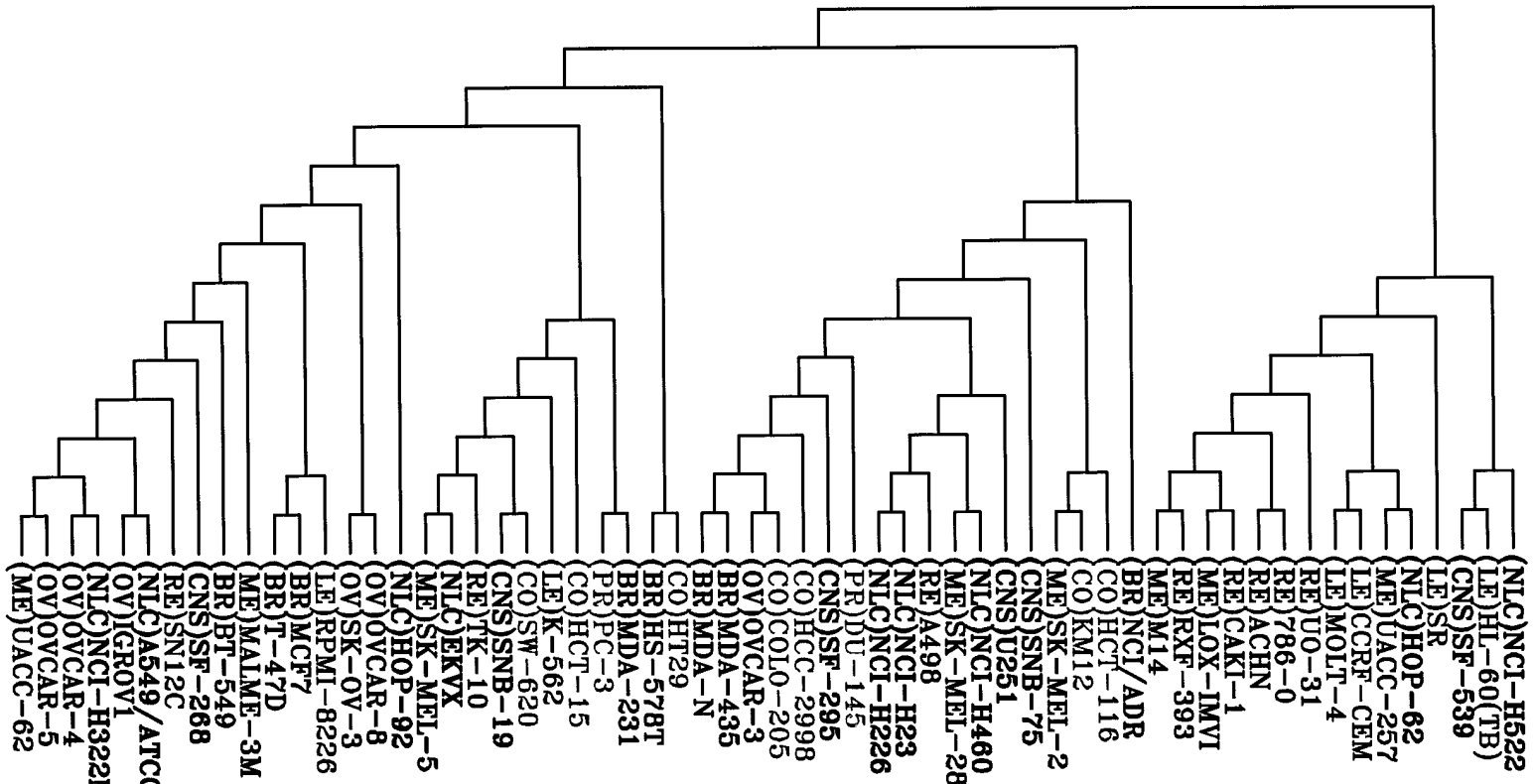
GROUP 25

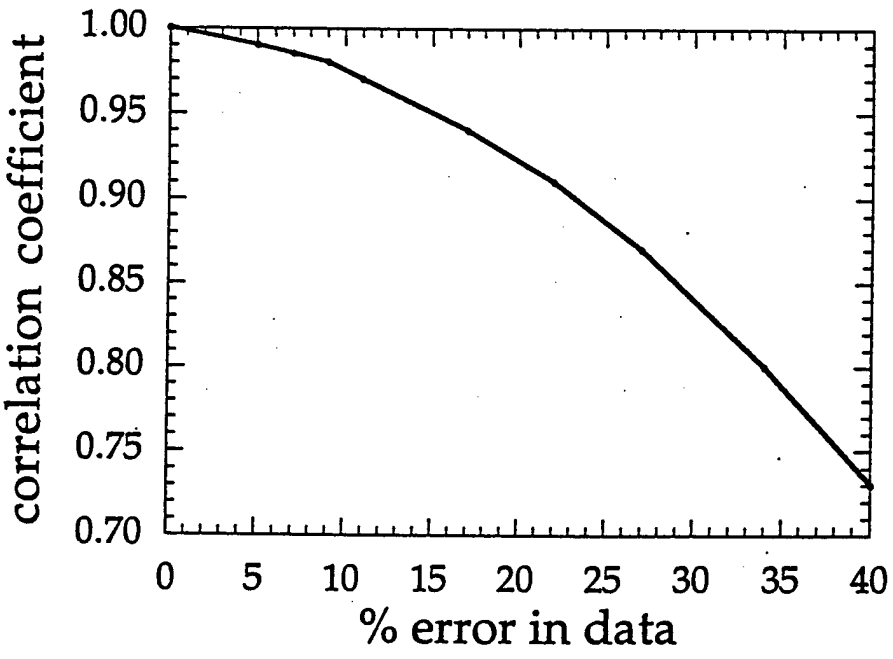

NLC-H23
NLC-H522
NLC-A549
NLC-EKVX
NLC-H226
NLC-H322M
NLC-H460
NLC-HOP62
NLC-HOP92
CO-HT29
CO-HCC2998
CO-HCT116
CO-SW620
CO-COLO205
CO-HCT15
CO-KM12
BR-MCF7
BR-ADR
BR-MDA231
BR-HS578T
BR-MDA435
BR-MDA-N
BR-BT549
BR-T47D
OV-CAR3
OV-CAR4
OV-CAR5
OV-CAR8
OV-IGROV1
OV-SKOV3
LE-CCRFCEM
LE-K562
LE-MOLT4
LE-HL60(TB)
LE-RPMI8226
LE-SR
RE-UO31
RE-SN12C
RE-A498
RE-CAKI1
RE-RXF393
RE-7860
RE-ACHN
RE-TK10
ME-LOXIMVI
ME-MALME3M
ME-SKMEL2
ME-SKMEL28
ME-SKMEL5
ME-M14
ME-UACC257
ME-UACC62
PR-PC3
PR-DU145
CNS-SNB19
CNS-SNB75
CNS-U251
CNS-SF268
CNS-SF295
CNS-SF539
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
CELL TYPE
CLUSTER NUMBER

60 Tumor Cell Panel

Clustered by Response to 122 Standard Agents

(Blk:NLC, Lt. Grn:CO, Blu:BR, Mag:LE, Red:OV, Dk. Green:RE, Brn:ME, Lt. Blu:PR, Blk:CNS)




correlation coefficient
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
5
10
15
20
25
30
35
40
% error in data

References

1. Gordon E. M.; Barrett R. W.; Dower W.; Fodor S. P.; Gallop M. A. Application of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies and Future Directions. *J. Med. Chem.* **1994**,*37*, 1385-1401.

2. Ganesan A. Combinatorial Chemistry in the Hunt for Medicines. *Nature* **1998**,*393*, 727.

3. Gray N. S.; Wodicka L.; Thunnissen A. M.; Norman T. C.; Kwon S.; Espinoza F. H.; Morgan D. O.; Barnes G.; Clerc S. L.; Meijer L.; Kim S. H.; Lockhart D. J.; Shultz P. G. Exploiting Chemical Libraries, Structure and Genomics in the Search for Kinase Inhibitors. *Science* **1998**,*281*, 533-538.

4. Boyd M. R. *The NCI in vitro Anticancer Drug Discovery Screen*. Totowa, New Jersey, Humana Press; 1995.

5. Kauver L. M. Affinity Fingerprinting. *Biotechnology* **1995**,*13*, 965-966.

6. Grever M. R.; Schepartz S. A.; Chabner B. A. The National Cancer Institute: Cancer Drug Discovery and Development Program. *Semin. Oncol.* **1992**,*19*, 622-638.

7. Monks A.; Scudiero D.; Skehan P.; Shoemaker R.; Paull K.; Vistica D.; Hose C.; Langely C.; Cronise P.; Vaigro-Wolff A.; Grey-Goodrich M.; Cambell L.; Mayo J.; Boyd M. R. Feasibility of a High Flux Anticancer Drug Screen Using a Diverse Panel of Cultured Human Tumor Cell Lines. *J. Natl. Canc. Inst.* **1991**,*83*, 757-766.

8. Chee M.; Yang R.; Hubbell E.; Berno A.; Huang X. C.; Stern D.; Winkler J.; Lockhart D. J.; Morris M. S.; Fodor S. P. Assessing Genetic Information with High-Density DNA Arrays. *Science* **1996**,*274*, 610-614.

9. Botstein D.;Cherry J. M. Molecular Linguistics: Extracting Information from Gene and Protein Sequences. *Proc. Natl. Acad. Sci. USA* **1997**,*94*, 5506-5507.

10. Zhang L.; Zhou W.; Velculescu V. E.; Kern S. E.; Hruban R. H.; Hamilton S. R.; Volgelstein B.; Kinzler K. W. Gene Expression Profiles in Normal and Cancer Cells. *Science* **1997**,*276*, 1268-1272.

11. Castell J. V.; Gomes-Lechon M. J. *In vitro Methods in Pharmaceutical Research*. San Diego, Academic Press; 1997.

12. Bellenson J. Integrating Information Technology and Drug Discovery processes. *Nature Biotechnol.* **1998**,*16*, 597-598.

13. Ajay A.; Walters W. P.; Murcko M. A. Can We Learn to Distinguish between Drug-like and Nondrug-like Molecules? *J. Med. Chem.* **1998**,*41*, 3314-3324.

14. Sadowski J.;Kubinyi H. A Scoring Scheme for Discriminating between Drugs and Non-drugs. *J. Med. Chem.* **1998**,*41*, 3325-3329.

15. Marchington T. From Data to Drugs. *Biotechnology* **1995**,*13*, 239-242.

16. Shi L. M.; Fan Y.; Myers T. G.; O'Connor P. M.; Paull K. D.; Friend S. H.; Weinstein J. N. Mining the NCI Anticancer Drug Discovery Databases: Genetic Function Approximation for the QSAR Study of Anticancer Ellipticine Analogues. *J. Chem. Inf. Comput. Sci.* **1998**,*38*, 189-199.

17. Gillet V. J.; Willett P.; Bradshaw T. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**,*38*, 165-179.

18. Benton D. Integrated Access to Genomic and Other Bioinformation: An Essential Ingredient of the Drug Discovery Process. *SAR QSAR Environ. Res.* **1998**,*8*, 121-155.

19. Shi L. M.; Myers T. G.; Fan Y.; O'Connor P. M.; Paull K. D.; Friend S. H.; Weinstein J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: Cluster Analysis of Ellipticine Analogs with p53-inverse and Central Nervous System Selective Patterns of Activity. *Mol. Pharmacol.* **1998**,*53*, 241-251.

20. O'Connor P. M.; Jackman J.; Bae I.; Myers T. G.; Fan S.; Mutoh M.; Scudiero D. A.; Monks A.; Sausville E. A.; Weinstein J. N.; Friend S.; Fornace A. J. J.; Kohn K. W. Characterization of the p53 Tumor Suppressor Pathway in cell Lines of the National Cancer Institute Anticancer Drug Screen and Correlations with the Growth-inhibitory Potency of 123 Anticancer Agents. *Cancer Res.* **1997**,*57*, 4285-4300.

21. Myers T. G.; Weinstein J. N.; O'Connor P. M.; Friend S. H.; Fornace A. J.; Kohn K. W.; Fojo T.; Bates S. E.; Rubenstein L. V.; Anderson N. L.; Buolamwini J. K.; Osdol W. W. V.; Monks A.; Scudiero D. A.; Sausville E. A.; Zaharevitz D. W.; Bunow B. B.; Viswanadhan V. N.; Johnson G. S.; Wittes R. E.; Paull K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**,*275*, 343-349.

22. Paull K. D.; Shoemaker R. H.; Hodes L.; Monks A.; Scudiero D. A.; in L. R.; Plowman J.; Boyd M. R. Display and Analysis of Patterns of Differential Activity of Drugs Against Human Tumor Cell Lines: Development of Mean Graph and COMPARE Algorithm. *J. Natl. Cancer. Inst.* **1989**,*81*, 1088-1092.

23. Paull K.; Hamel E.; Malspeis L. *Prediction of Biochemical Mechanism of Action from the in vitro Antitumor Screen of the National Cancer Institute.* In Cancer Chemotherapeutic Agents (Ed. W. O. Foye) Washington D.C., ACS; 1995.

24. Boyd M.; Paull K. D. Some Practical Considerations and Applications of the National Cancer Institute in vitro Anticancer Drug Discovery Screen. *Drug Devel. Res.* **1995**,*34*, 91-109.

25. Hrach K. Comparison of Survival Between Two Groups Using Software SAS, S-PLUS and STATISTICA. *Int. J. Med. Inf.* **1997**,*45*, 31-33.

26. Weinstein J. N.; Kohn K. W.; Grever M. R.; Viswanadhan V. N.; Rubinstein L. V.; Monks A. P.; Scudiero D. A.; Welch L.; Koutsoukos A. D.; Chiausa A. J.; Paull K. D. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**,*258*, 447-451.

27. Koutsoukos A. D.; Rubenstein L. V.; Faraggi D.; Simon R. M.; Kalyandrug S.; Weinstein J. N.; Kohn K. W.; Paull K. D. Discrimination Techniques Applied to

the NCI in vitro Anti-tumor Drug Screen: Predicting Biochemical Mechanism of Action. *Stat. Med.* **1994**,*13*, 719-730.

28. Shi L. M.; Fan Y.; Myers T. G.; Waltham M.; Paull K. D.; Weinstein J. N. Mining the Anticancer Activity Database Generated by the U.S. National Cancer Institute's Drug Discovery Program Using Statistical and Artificial Intelligence Techniques. **In** *Modeling and Scientific Computing* **1997**,

29. vanOsdol W. W.; Myers T. G.; Paull K. D.; Kohn K. W.; Weinstein J. N. Use of Kohonen Self-organizing Map to Study the Mechanism of Action of Chemotherapeutic agents. *J. Natl. Cancer Inst.* **1994**,*86*, 1853-1859.

30. Maggiora G.; Johnson M. A. *Concepts and Applications of Molecular Similarity*. NY, John Wiley; 1990.

31. Janin J.;Chothia C. The Structure of Protein-protein Recognition Sites. *J. Biol. Chem.* **1990**,*265*, 16027-16030.

32. Clackson T.;Wells J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science* **1995**,*267*, 383-386.

33. Schreiber G.; Fersht A. R. Energetics of Protein-Protein Interactions: Analysis of the Barnase-Barstar Interface by Single Mutations And Double Mutant Cycles. *J. Mol. Biol.* **1995**,*248*, 478-486.

34. Covell D. G.; Wallqvist A.; Rice W.; Turpin J. Information-Intensive Approaches to Drug Discovery: Identification and Testing of HIV-1 Integrase Inhibitors. **1997**,*in preparation.*

35. Harary F. *Graph Theory*. Reading, MA, Addison-Wesley; 1971.

36. Berry M. W.; Dumais S. T.; O'Brien G. W. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Rev.* **1995**,*37*, 573-595.

37. Liu K. Application of SVD in Optimization of Structural Modal Test. *Computers and Structures* **1997**,*63*, 51-59.

38. Golub G.; Loan C. V. *Matrix Computations*. Baltimore, MD, Johns-Hopkins University Press; 1989.

39. Bahar I.; Wallqvist A.; Covell D. G.; Jernigan R. L. Correlation Between Native State Hydrogen Exchange Data and Cooperative Residue Fluctuations from a Simple Model. *Biochemistry* **1998**,*37*, 1067-1075.

40. Sneath P. H. A.;Sokal R. R. *Numerical Taxonomy*. San Francisco, W. H. Freeman and Company; 1973.

41. Giuliani A.; Colosimo A.; Benigni R.; Zbilut J. On the Constructive Role of Noise in Spatial Systems. *Phys. Lett. A* **1998**,*247*, 47-52.

42. Bahar I.; Atilgan A. R.; Jernigan R. L.; Erman B. Understanding the Recognition of Protein Structural Classes by Amino Acid Composition. *Proteins* **1997**,*29*, 172-185.

43. Chabner B. A.; Longo D. L. *Cancer Chemotherapy and Biotherapy: Principles and Practice*. Philadelphia and New York, Lippencot-Raven; 1996.

44. Pratt W. B.; Ruddon R. W.; Ensminger W. D.; Maybaum J. *The Anticancer Drugs*. New York, Oxford University Press; 1994.

45. Mardia K. V.; Kent J. T.; Biby J. M. *Multivariate Analysis*. London, Academic Press; 1979.

46. Hwang J.; Shyy S.; Chen A. Y.; Juan C. C.; Whang-Peng J. Studies of Topoisomerase-Specific Antitumor Drugs In Human Lymphocytes Using Rabbit Antisera Against Recombinant Human Topoisomerase II Polypeptide. *Cancer Res.* **1989**,*49*, 958-962.

47. Bernstein F. C.; Koetzle T. F.; Williams G. J. B.; Jr. E. F. M.; Brice M. D.; Rogers J. R.; Kennard O.; Shimanouchi T.; Tasumi M. The Protein Data Bank: A Computer Based Archival file for Macromolecular Structures. *J. Mol. Biol.* **1977**,*112*, 535-542.

48. Stryer L. *Biochemistry*. New York, W. H. Freeman and Company; 1988.

49. Nogales E.; Wolf S. G.; Downing K. H. Structure of the α/β Tubulin Dimer by Electron Crystallography. *Nature* **1998**,*391*, 199-203.

50. Nogales E.; Downing K. H.; Amos L. A.; Lowe J. Tubulin and FtsZ Form a Distinct Family of GTPases. *Nature Struct. Biol.* **1998**,*5*, 451-458.

51. Nogales E.; Whittaker M.; Milligan R. A.; Downing K. H. High Resolution Model of the Microtubule. *Cell* **1999**,*96*, 79-88.

52. Martin Y. C.; Willet P. *Designing Bioactive Molecules*. Washington D. C, ACS; 1998.

53. Cummins D. J.; Andrews C. W.; Bentley J. A.; Cory M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Comp. Inf. Comput. Sci.* **1996**,*36*, 750-763.

54. Shemetulskis N. E.; Jr. J. B. D.; B.W. Dunbar; Moreland D. W.; Humblet C. Enhancing the Diversity of Corporate Databases Using Chemical Database Clustering and Analysis. *J. Comput. Aided Mol. Des.* **1995**,*9*, 407-416.

55. Lewis R. A.; Mason J. S.; McLay I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: the Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**,*37*, 599-614.

56. Good A. C.;Lewis R. A. New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning up the Design Process with HARpick. *J. Med. Chem.* **1997**,*40*, 3926-3936.

57. Weininger D.; Weininger A.; Weininger J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notations of Combinatorial Libraries: the Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**,*37*, 599-614.

58. Brown R. D.;Martin Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1996**,*37*, 1-9.

59. Burden F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**,*29*, 225-227.

60. Randic M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**,*37*, 672-687.

61. Pearlman R. S.; Smith K. M. Novel Software Tools for Chemical Diversity. *Perspectives in drug discovery* **1998**,339-353.

62. Miller M. D. SQ - *A Program for Producing Rapid Molecular Superimpositions*. ACS 210th National Meeting, **1995** Abstract.

63. Snedecor G. W.; Cochran. W. G. *Statistical Method*. Ames, Iowa, Iowa State University Press; **1980**.

64. Alvarez M.; Paull K.; Monks A.; Hose C.; Lee J. S.; Weinstein J.; Grever M.; Bates S.; Fojo T. Generation of a Drug Resistance Profile by Quantitation of MDR-1/P-glycoprotein in the Cell Lines of the National Cancer Institute Anticancer Drug Screen. *J. Clin. Invest.* **1995**,*95*, 2205-2214.

65. Lee J. S.; Paull K.; Alvarez M.; Hose C.; Monks A.; Grever M.; Fojo A. T.; Bates S. Rhodamine Efflux Patterns Predict P-glycoprotein Substrates in the National Cancer Institute Drug Screen. *Mol. Pharmacol.* **1994**,*46*, 627-638.

66. Wu L., Smythe A.M., Stinson S.F., Mullendore L.A., Monks A., Scuderio D.A., Paull K.D., Koutsoukos A.D., Rubinstein L.V., Boyd M.R. and Shoemaker R.H. Multidrug-resistant Phenotype of Disease-oriented Panels of Human Tumor Cell Lines Used for Anticancer Screening. Cancer Res. **1992**, 52:3029-3034.

Appendix A. Survey results from an analysis of available crystal structures complexed with ligands that are structurally similar to the standard anticancer agents analyzed here.

Table 4 lists the protein complexes identified here for investigating this issue. Our intention here is not to provide a complete list of all structural analogs within the Protein Data Bank (PDB) [47], but to indicate the range of protein structures that are known to form complexes with the structural analogs to the 122 anticancer agents. The results presented in Table 4 were obtained using the SMILES-based searching tools available in the RELIBASE part of the PDB browser (http://www.pdb.bnl.gov). The first column in the table describes the types of enzymes, the second and third give the name and PDB identifier of each enzyme, the fourth column is the ligand bound in the complex, and the fifth column lists the anticancer agents that are either identical or structural analogs to the standard 122 anticancer agents.

The results in Table 4 directly indicate the sites of action of many of the agents assigned to groups 9-25 of our cluster analysis. For example, crystallographic complexes exist for most of the enzymes involved in pyrimidine biosynthesis pathway. This pathway involves six enzymatically catalyzed steps. The CAD gene encodes a trifunctional protein associated with the activity of the first three enzymes in this six-step pathway: carbamoylphosphate synthase (EC 6.3.5.5), aspartate transcarbamoylase (EC 2.1.3.2), and dihydroorotase (EC 3.5.2.3) -also referred to as CPSase, ATCase and DHOase, respectively-. Crystallographic complexes exist for acivicin (163501) bound to CPSase, PALA (224131) bound to ATCase and brequinar (368390) bound to DHOase. In addition, the sites of action of methotrexate (740) as well as other folate byproducts include dihydrofolate reductase, thymidylate synthase, AICAR transformylase and GAR transformylase; all of which are included in the set of complexes listed in Table 4. Purine biosynthesis occurs by *de novo* pathways as well as from preformed nucleosides and nucleotides via salvage reactions [48]. Phosphoribosyl kinases and transferases are involved in both processes, and are found in crystallographic complex with many of the nucleoside analogs included in this study. A surprising finding includes the recent dimeric structure of tubulin in complex with a taxane. A nucleoside analog is also bound at the dimer interface between the α and β tubulin subunits [49-51]. Taken together, these crystallographic complexes indicate that many of the

antitumor agents included in these groups target one or in some cases many proteins involved in nucleic acid biosynthesis or mitosis. The cell screening patterns of these compounds, when clustered according to the methods used here, clearly separate these compounds from DNA-damaging agents.

Table 4. Proteins Complexed with Ligands Similar to Anticancer Agents.

Enzyme Class	Name	PDB	ligand	NSC
ligase	carbamoyl phosphate synthase	1jdb	GLN chan	163501(D)
	,,	1jdb	ADP	71851,71261(D)
hydrolase	cytidine deaminase	1aln	3-deazacytidine	102816(D),143095(G)
	,,	1ctt	dihydrozebularine	102816(D),143095(G, 264880(E)
	,,	1ctu	zebularine	148958(D),264880(E)
oxido-reductase	dihydroorotate dehydrogenase	2dor	flavin mononucleotide	148958(D),2764 0(G)
	,,	2dor	orotic acid	148958(D)
	diaminopimelic acid dehydrogenase	1dap	NDP	71851,71261(D)
	,,	1DAP	DA3	163501(D)
	cyclooxygenase	3pgh	flurbiprofen	368390(E)
	dihydrofolate reductase	1ai9	NDP	71851,71261(D)
	,,	1ao8	MTX	740(G)
	,,	1dhf	MTX	740(G)
transferase	thymidylate synthase	1bjg	5-F-deoxyuridine	148958(G)
	,,	1bjg	hydrofolic acid	623017,174121(G)
	,,	1vzd	dideazafolic acid	134033(G)
	,,	2tdd	hydrofolic acid	134033(G)
	,,	1tls	5-F-deoxyuridine	148958(G)
	,,	1lce	hydrofolic acid	132483(G)
	amidotransferase carbamoyl phosphate synthetase	1a9x	GLN	163501(D)
	,,	1a9x	ADP	71851,71261(D)
	,,	2tdd	hydrofolic acid	134033(G)
	,,	1tls	5-F-deoxyuridine	148958(G)
	,,	1lce	hydrofolic acid	132483(G)
	,,	1a9x	GLN	163501(D)
	aspartate transcarbamylase	1acm	PALA	224131(D)
	phosphoribosyl transferase	1opr	orotic acid	148958,102816(D)
	,,	1sto	orotidine	148958,27640(D)
	carbamoyl transferase	1rai	cytidine	102816,27640(D)
	phosphoribosyglycinamide	1cde	ribonucleotide	102816(D)
	formyltransferase	1gar	U89	118994,71851 71261(D)
	methyltransferase	1v39	homocysteine	71261,71851(D)

transferase	nucleotidyl transferase	1waf	GMP	71261,71851(D)
	thioredoxin	1t7p	guanosine	71261,71851(D)
	nucleoside phosphorylase	1a69	formycin	143095(G)
	,,	1a9t	hypoxanthine	71851,71261(D)
	,,	1a9t	ribose-1-phosphate	102816(D)
	diphosphate kinase	1be4	guanosine	71261,71851(D)
	diphosphate kinase	1kdn	ADP	71261,71851(D)
	adenylate kinase	1dvr	adenosine	71261,71851(D)
	thymidine kinase	1kim	thymidine	27640(G)
	protein kinase inhibitor	1kpe	adenosine	71261,71851(D)
	purine phosphorylase	1vfn	hypoxanthine	71851, 71261(D)
	UMP/CMP kinase	2ukd	ADP C5P	71851, 71261(D)
Microtubules	α/β tubulin dimer	1tub	gtp,gdp	71851, 71261(D)
		1tub	taxotere	125973